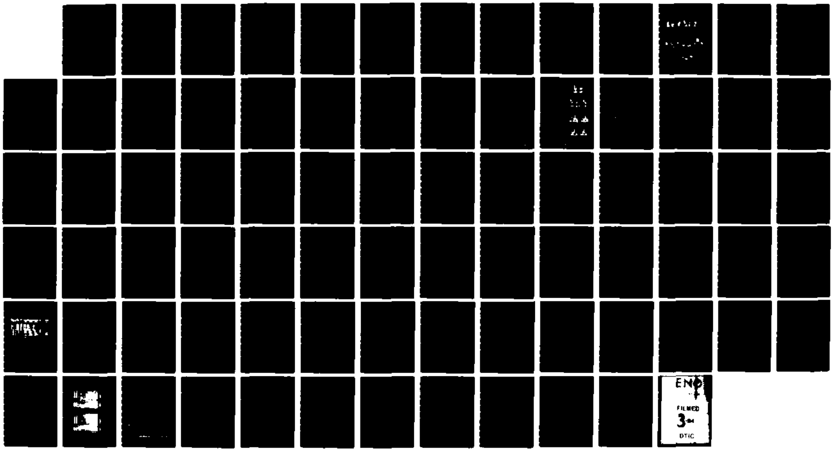
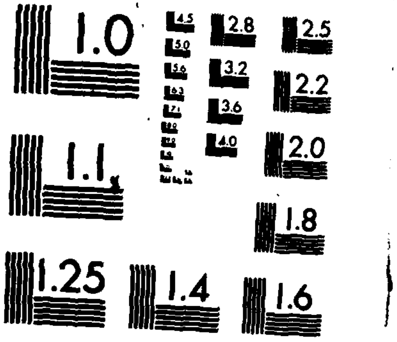
AD-A137 697 SPEECH RECOGNITION: ACOUSTIC PHONETIC AND LEXICAL KNOWLEDGE REPRESENTATION(U) MASSACHUSETTS INST OF TECH CAMBRIDGE RESEARCH LAB OF ELECTRONICS V W ZUE 1/1

UNCLASSIFIED 01 FEB 84 N00014-82-K-0727 F/G 5/7 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

STATUS REPORT

SPEECH RECOGNITION: ACOUSTIC PHONETIC & LEXICAL KNOWLEDGE REPRESENTATION

OFFICE OF NAVAL RESEARCH
DEPARTMENT OF THE NAVY

Contract Number N00014-82-K-0727

covering the periods
July 1, 1983 - September 30, 1983
October 1, 1983 - December 31, 1983

Submitted by:
Victor W. Zue
February 1, 1984

Massachusetts Institue of Technology
Research Laboratory of Electronics
Cambridge, Massachusetts 02139

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. AD-A137 697	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) SPEECH RECOGNITION: ACOUSTIC PHONETIC AND LEXICAL KNOWLEDGE REPRESENTATION		5. TYPE OF REPORT & PERIOD COVERED Status Report 07/01/83 - 09/30/83 and 10/01/83 - 12/31/83
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Victor W. Zue		8. CONTRACT OR GRANT NUMBER(s) N00014-82-K-0727
9. PERFORMING ORGANIZATION NAME AND ADDRESS Research Laboratory of Electronics Massachusetts Institute of Technology Cambridge, Massachusetts 02139		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 049-542
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research - Department of the Navy - 800 North Quincy Street Arlington, Virginia 22217		12. REPORT DATE February 1, 1984
		13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

A large vocabulary and a continuous speech recognition system, which are acoustic-phonetically based phontactic constraints acoustic cues to word boundaries.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The purpose of this program is to develop a speech data base facility under which the acoustic characteristics of speech sounds in various contexts can be studied conveniently; investigate the phonological properties of a large lexicon of, say 10,000 words, and determine to what extent the phonotactic constraints can be utilized in speech recognition; study the acoustic cues that are used to mark work boundaries; develop a test bed in the form of a large-vocabulary, IWR system to study the interactions of acoustic, phonetic and lexical knowledge; and develop a limited continuous speech recognition system with the goal of recognizing any English word from its spelling in order to assess the interactions of higher-level knowledge sources.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TECHNICAL PROGRESS
Contract Number N00014-82-K-0727
Periods covered: 07-01-83 through 09-30-83 and 10-01-83 through 12-31-83

Over the past several months we have broadened the scope of our investigation into the lexical and phonetic constraints imposed by the English language. At the segmental level. we investigated the effects of segmentation and classification uncertainties. both of which are likely to occur in the actual implementation of a phonetic front-end. The results of our experiments indicate that. if the uncertainties and errors are reasonable. the constraints can still be very powerful. We have also conducted a number of experiments examining the functional loads carried by segments in stressed versus unstressed syllables. We found that the stressed syllables provide a significantly greater amount of constraining power than unstressed syllables. This implies that in the actual implementation, acoustic-phonetic information around unstressed syllable should not receive undue emphasis. At the prosodic level. we started to investigate the constraints imposed by the stress pattern of words. Preliminary results indicate that knowledge about the stress pattern, or simply the relative position of the syllable with primary stress, greatly constrains the number of word candidates.

Implementation of the large-vocabulary, isolated-word recognition system has progressed in several directions. First, the performance of the acoustic classifier was evaluated on the speech data from two speakers, one male and one female. After minor modifications, we feel that its performance, including edge detection and parameter characterization, is quite satisfactory, although some of the rules are still not adequate. Second, a software system, called TRANSCRIBE, has been written. This is an interactive system that allows researchers to write acoustic-phonetic rules and evaluate their performance on a database. The system has the capability of explaining the history of how a rule has been triggered, as well as why certain rules failed to apply. This facility, together with the speech data that we have collected previously, has greatly improved our ability to specify and debug acoustic-phonetic rules. As a consequence, we expect that we will be able to complete the broad acoustic-phonetic classifier within the next several months. Third, we have implemented a system that locates the stressed and reduced syllables of a word. Thus. for example. the system can determine that the first syllable of the word "institute" is stressed, whereas the second syllable is reduced.

The continuous digit recognition system has been implemented up to the level of lexical access. and we have just completed our first round of evaluation. The system was developed using the speech data from one male speaker, and the initial evaluation was performed using some two hundred digits spoken by three new speakers. one male and two female. We found that the error rate is approximately 1%. That is. the correct digit is not one of the candidates in less than 1% of the time. The corresponding depth of the digit lattice is approximately 3. While these results are very preliminary. we are nevertheless encouraged and feel that this may be a viable approach to speaker-independent digit recognition. We are continuously refining the system. and we expect to make another performance evaluation in the near future. this time over a larger database. In addition. we have also started to collect data on other languages (Japanese. French. and Italian) for the digit task. Digits strings for these languages have been recorded and digitized. Spectrograms were made. and the acoustic-phonetic rules appropriate for the language in question are specified both within a digit and across digit boundaries. We expect to evaluate the performance of the basic digit recognition system for different languages. thus determining the effectiveness of the rules. in the next quarter.

We are continuing our effort to find cues to delineate words in continuous speech. It is well known that words in continuous speech are not separated by pauses. In some cases. the acoustic characteristics can be significantly different. depending on the location of the word boundary. Thus. for example. the acoustic properties of phrases "nitrate" and "night rate" may be quite different. Before investigating the possible acoustic differences between such phrases. we first investigated the distributional constraints imposed by the English language. We asked the following question: Given a consonant sequence. can one determine whether this sequence can only occur at word boundaries? Using text files ranging from 200 to 38,000 word, we found that, one the average. 80% of consonant sequences found can only occur at word boundaries. In other words, only one out of five consonant sequences can occur word internally as well as across word boundaries. Thus given an ideal phonetic transcription. the word boundary can be determined uniquely most of the time. Further studies along this line will continue in the next quarter.

The alignment of a speech signal with its corresponding phonetic transcription is an essential process in speech research. since the time-aligned transcription provides direct access to specific phonetic events in the signal. Traditionally. the alignment is done manually by a trained acoustic phonetician. The task. however. is prone to error. tedious and extremely time consuming. During the past six month we initiated an effort to develop a system that performs the time-alignment automatically. The alignment is achieved using a standard pattern classification algorithm and a dynamic programming algorithm, augmented with acoustic-phonetic constraints. In initial implementation of the system has been completed. We will refine some of its components and perform formal evaluation during the next quarter.

Extensive support software of SPIRE. SPIREX. and LEXIS has been written. These three programs have now been field tested in a number of research laboratories around the country. including several defense contractors and installations. With the arrival of the Symbolics-3600 Lisp machine. programs have been converted such that they now run on all the Lisp machines. We continue to increase the amount of speech data. We now have more than one hour of digitized speech available on line.

PUBLICATIONS AND PRESENTATIONS

:

The following papers, describing work supported at least in part by this contact. have been presented at various professional meetings: (copies of the papers are included with this Report)

Zue, Victor W., (1983) "The Use of Phonetic Rules in Automatic Speech Recognition," Invited Paper, 11*th* International Congress on Acoustics, July 15-16, Toulouse, France.

Zue, Victor W., (1983) "Proposal For An Isolated-Word Recognition System Based On Phonetic Knowledge and Structural Constraints." Invited Paper, special session on "Human and Automatic Speech Recognition" at the Tenth International Congress of Phonetic Science, August 1-6, Utrecht, the Netherlands.

Huttenlocher, Daniel P. and Zue, Victor W., (1983) "Phonotactic and Lexical Constraints in Speech Recognition," paper presented at the American Association for Artificial Intelligence Annual Conference, August 21-26, Washington, D. C.

Huttenlocher, Daniel P. and Zue, Victor W., (1983) "Exploring Phonotactic and Lexical Constraints in Word Recognition," paper presented at the 106th Meeting of the Acoustical Society of America, November 8, San Diego, CA.

Chen, Francine R. and Zue, Victor W., (1983) "Exploring Allophonic and Lexical Constraints in a Continuous Speech Recognition System," paper presented at the 106th Meeting of the Acoustical Society of America, November 8, San Diego, CA.

Lamel, Lori F., (1983) "The Use of Structural Constraints to Determine Word Boundaries," paper presented at the 106th Meeting of the Acoustical Society of America, November 8, San Diego, CA.

Zue, Victor W., (1983) "Speech Recognition Research at MIT," paper presented at the Department of Defense, Voice SubTAG Meeting, October 25-26, Fort Monmouth, New Jersey.

Zue, Victor W. and Huttenlocher, Daniel P., (1983) "Computer Recognition of Isolated Words From Large Vocabularies: Lexical Access Using Partial Phonetic

Information," paper presented at the International Conference on Advanced Automation, December, Taiwan, Republic of China.

The following papers, describing work supported at least in part by this contract, have been accepted for publication at various professional meetings:

Chen, Francine R. and Zue, Victor W., (1984) "Application of Allophonic and Lexical Constraints in Continuous Digit Recognition," paper to be presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, March 19-21, San Diego, CA.

Huttenlocher, Daniel P. and Zue, Victor W., (1984) "A Model of Lexical Access from Partial Phonetic Information," paper to be presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, March 19-21, San Diego, CA.

Lamel, Lori F. and Zue, Victor W., (1984) "Properties of Consonant Sequences within Words and Across Word Boundaries," paper to be presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, March 19-21, San Diego, CA.

Leung, Hong C. and Zue, Victor W., (1984) "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech," paper to be presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, March 19-21, San Diego, CA.

Presented at the 11th ICA, Toulouse Symposium, July, 1983

Speech Communication 2 (1983) 181–186
North-Holland

INVITED LECTURE

THE USE OF PHONETIC RULES IN AUTOMATIC SPEECH RECOGNITION

Victor W. ZUE

Department of Electrical Engineering & Computer Science and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received 8 April 1983

1. Introduction

Automatic speech recognition by machine is a research topic that has fascinated many speech scientists for more than forty years. For many, it represents the ultimate challenge to our understanding of the production and perception processes of human communication. However, the last decade has witnessed a flourish of research efforts in the development of speaker-dependent, small-vocabulary, isolated-word recognition systems that utilize little or no speech-specific knowledge. These systems derive their power primarily from general-purpose pattern recognition techniques. While these techniques are adequate for a small class of well-constrained speech recognition problems, their extendibility to tasks involving multiple speakers, larger vocabularies and/or continuous speech is questionable.

Reliance on general pattern matching techniques has been partly motivated by the unsatisfactory performance of early phonetically-based speech recognition systems. The difficulty of automatic acoustic-phonetic analysis has also led to the speculation that phonetic information must be derived primarily from semantic, syntactic and discourse constraints rather than from the acoustic signal. The poor performance of the early phonetically-based systems can be attributed mainly to our limited knowledge of the context-dependency of the acoustic characteristics of speech sounds. However, this picture is slowly changing. We now have a far better understanding of contextual influences on phonetic segments. This improved understanding has been demonstrated in a series of spectrogram reading experiments (Zue and Cole, 1979; Cole et al., 1980; Cole and Zue 1980). It was found that a trained subject can phonetically transcribe unknown sentences from speech spectrograms with an accuracy of approximately 85%. This performance is better than the phonetic recognizers reported in the literature, both in accuracy and rank order statistics. It was also demonstrated that the process of spectrogram reading makes use of explicit acoustic phonetic rules, and that this skill can be learned by others. These results suggest that the acoustic signal is rich in phonetic information, which should permit substantially better performance in automatic phonetic recognition.

One of the most important factors contributing to the good performance of spectrogram reading is our improved understanding of the acoustic characteristics of fluent speech. To be sure, there has been ongoing research on the acoustic properties of speech sounds over the past few decades, and a great deal of knowledge has been acquired. However, with few exceptions, these research efforts have been focused on the acoustic properties of consonants and vowels in stressed consonant-vowel syllables. It was not until the past decade that researchers began to focus on the acoustic characteristics of speech sounds in continuous speech. We now have a much better understanding of the properties of speech sounds in different phonetic environments [see, for example, Umeda (1975), Kameny (1975), Klatt (1975), Zue (1976), Umeda (1977)]. Furthermore, we are beginning to develop a quantitative understanding of the phonological processes governing the concatenation of words [see, for example, Oshika et al. (1975), Cohen and Mercer (1975)]. A few of the effects have even been studied in detail (Zue and Laferriere, 1979; Zue and Shattuck-Hufnagel, 1980). In addition, as

0167-6393/83/$3.00 © 1983. Elsevier Science Publishers B.V. (North-Holland)

a consequence of studies on the properties of speech sounds and of the auditory responses to speech-like sounds [see, for example, Kiang (1980)], we are gaining better insight into how the speech signal is processed in the auditory system, what portions of the signal carry the principal information concerning distinctive phonetic dimensions, and what portions show more variability with respect to these dimensions. For example, the role of the burst spectra, burst amplitudes, and rapid onsets and offsets in identifying place of articulation and other features for stop consonants have been documented (Zue, 1976; Blumstein and Stevens, 1979).

In summary, we must emphasize that our ability to extract a great deal of phonetic information from the acoustic signal is primarily a reflection of our improved understanding of the factors that contribute to the phonetic identities of speech sounds and their acoustic correlates. Spectrogram reading is nothing more than a *paradigm* to demonstrate how the acoustic cues for phonetic contrasts are encoded in the speech signal. Native speakers of a language demonstrate this ability whenever they communicate by voice. In the remainder of this paper, we will discuss how phonetic information is encoded in the speech signal. We will also present some alternative ways to represent such information in speech recognition systems.

2. Phonetic variability in the speech signal

Phonotactic constraints

The speech signal is the output of a highly constrained system. In addition to having a very limited inventory of possible phonemes, a given language is also constrained with regard to the ways in which these phonemes can combine to form meaningful words. Knowledge about such constraints is implicitly possessed by native speakers of a language. For example, a native English speaker knows that 'vnuk' is not an English word. He/she also knows that if an English word starts with three consonants, then the first consonant must be an /s/, and the second consonant must be a voiceless stop (i.e. either /p/, /t/, or /k/). Such phonotactic knowledge is presumably very useful in speech communication, since it provides native speakers with the ability to fill in phonetic details that are otherwise not available or are distorted. Thus, as an extreme example, a word such as 'splint' can be recognized without having to specify the detailed phonetic features of the phonemes. In fact, 'splint' is one of only two words in the Merriam Pocket Dictionary (containing about 20000 words) that satisfies the following description:

[consonant][consonant][liquid or glide]
[vowel][nasal][stop].

While the existence of phonotactic constraints is well known, a recent set of studies (Shipman and Zue, 1982; Huttenlocher and Zue, 1983) provides a glimpse of the magnitude of their predictive power. These studies examine the phonotactic constraints of American English from the phonemic distributions in the 20000-word Merriam Webster's Pocket Dictionary. In one study the phonemes of each word were mapped into one of six broad phonetic categories: vowels, stops, nasals, liquids and glides, strong fricatives, and weak fricatives. Thus, for example, the word 'speak', with a phonemic string given by /spik/, is represented as the pattern:

[strong fricative][stop][vowel][stop].

It was found that, even at this broad phonetic level, approximately $\frac{1}{3}$ of the words in a 20000-word lexicon can be uniquely specified. In general the size of the equivalence class, (namely, the number of words sharing the same pattern), was quite small. The average size of the equivalence classes for the 20000-word lexicon was found to be approximately 2, and the maximum size was approximately 200. In other words, in the worst case, a broad phonetic representation of the words in a large lexicon reduces the number of possible word candidates to about 1% of the lexicon. Furthermore, over half of the lexical items belong to equivalence classes of size 5 or less.

Allophonic and phonological variations

When speech sounds are connected to form larger linguistic units, the canonical acoustic characteristics of a given speech sound will change as a function of its immediate phonetic environment. As an illustrative example, consider the utterance, "Tom Burton tried to steal a butter plate," shown in Fig. 1. Every word, except 'a', in this sentence contains a single occurrence of the phoneme /t/. However, depending upon the immediate phonetic environment and stress pattern, the underlying /t/'s are realized alternatively as an aspirated /t/ ('Tom'), an unaspirated /t/ ('steal'), a retroflexed /t/ with extended aspiration ('tried'), an unreleased /t/ ('plate'), a flap ('butter'), or a glottal stop ('burton'). The acoustic characteristics of these realizations are seen to be drastically different.

The modification of the acoustic properties of speech sounds as a function of the phonetic environment is not a phenomenon that is restricted to be within a word. When words are concatenated to form phrases and sentences, significant acoustic changes can result, as evidenced in the following example. Fig. 2 shows a spectrogram of the seven words 'did', 'you', 'meet', 'her', 'on', 'this', and 'ship', spoken in isolation as well as in a sentence, "Did you meet her on this ship?" We can see, for example, that the word-final /d/ and the word-initial /y/ in the word pair 'did you' are realized acoustically as a single /ǰ/; the word-final /t/ and the word-initial /h/ in the word pair 'meet her' are realized as a single flap; and the word-final /s/ and the word-initial /š/ in the word pair 'this ship' are realized as a single, long /š/. Such phonetic changes at word boundaries, particularly when there are adjacent word-final and word-initial consonants, are extremely common in American English. In order to properly perform lexical access, the nature of these phonological rules must be understood.

3. Representation of phonetic knowledge

Even though the acoustic realizations of phonetic segments are highly context-sensitive, most of the variations, such as the ones illustrated in Figs. 1 and 2, are systematic and can be captured by explicit rules. (For example, /t/ becomes a glottal stop [?] when preceded by a stressed vowel and followed by a syllabic nasal [n̩], as in 'Burton'.) Over the past decade, research in fluent speech has enabled us to gain a good understanding of the nature of these rules and how they interact. Although our present knowledge of the inventory of these rules is still incomplete, such knowledge, however fragmented, must be incorporated into a speech recognition system so that words can be recognized from seemingly ambiguous acoustic signals.

In order to discuss how acoustic phonetic knowledge should be represented in speech recognition systems, it is perhaps useful to distinguish three types of phonetic/phonological rules. First, there are the *phonotactic* constraints governing the allowable combination of phonemes. For example, the homorganic rule in English specifies that a syllable-final nasal/stop cluster must agree in the place of articulation. It should be noted that the obligatory nature of the phonotactic constraints makes them more suited to categorical formulations.

Second, there are the rules that describe the modification of acoustic characteristics of phonemes in various phonetic environments. These *allophonic* rules are again mostly categorical. However, their acoustic consequences may take on a continuum of values. For example, in American English the phoneme /t/ becomes a retroflexed alveolar voiceless stop when preceding a retroflexed consonant or vowel. On the other hand, the amount of the acoustic change, such as the lowering of the burst frequency and the lengthening of the voice onset time, may vary over a wide range. Traditionally, allophonic variations have been considered one of the major sources of difficulty for speech recognition, since they represent undesirable distortion, or noise, imposed on the canonic characteristics of the phonemes. However, researchers are beginning to see such allophonic variations as a source of information [see, for example, Nakatani and Dukes (1977)]. For example, Church (1983) demonstrated that detailed knowledge about allophonic rules can be exploited during lexical access by parsing the phonetic string into syllables and other suprasegmental constituents.

Allophonic rules traditionally have been described in the context-dependent formalism: $A \rightarrow B/C\_D$ (i.e., segment *A* becomes segment *B* in the context of segments *C* and *D*). There are several

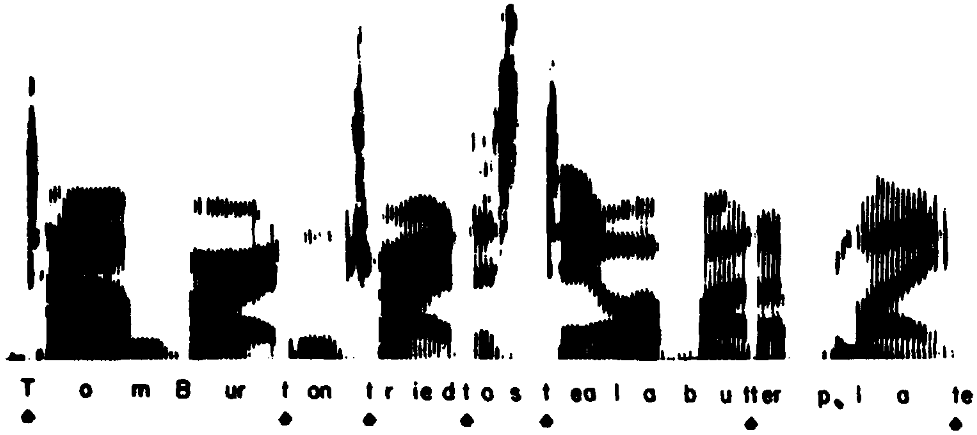


Fig. 1. Spectrogram of the sentences, "Tom Burton tried to steal a butter plate," spoken by a male speaker. The spectrogram illustrates the various acoustic realizations of the phoneme /t/.

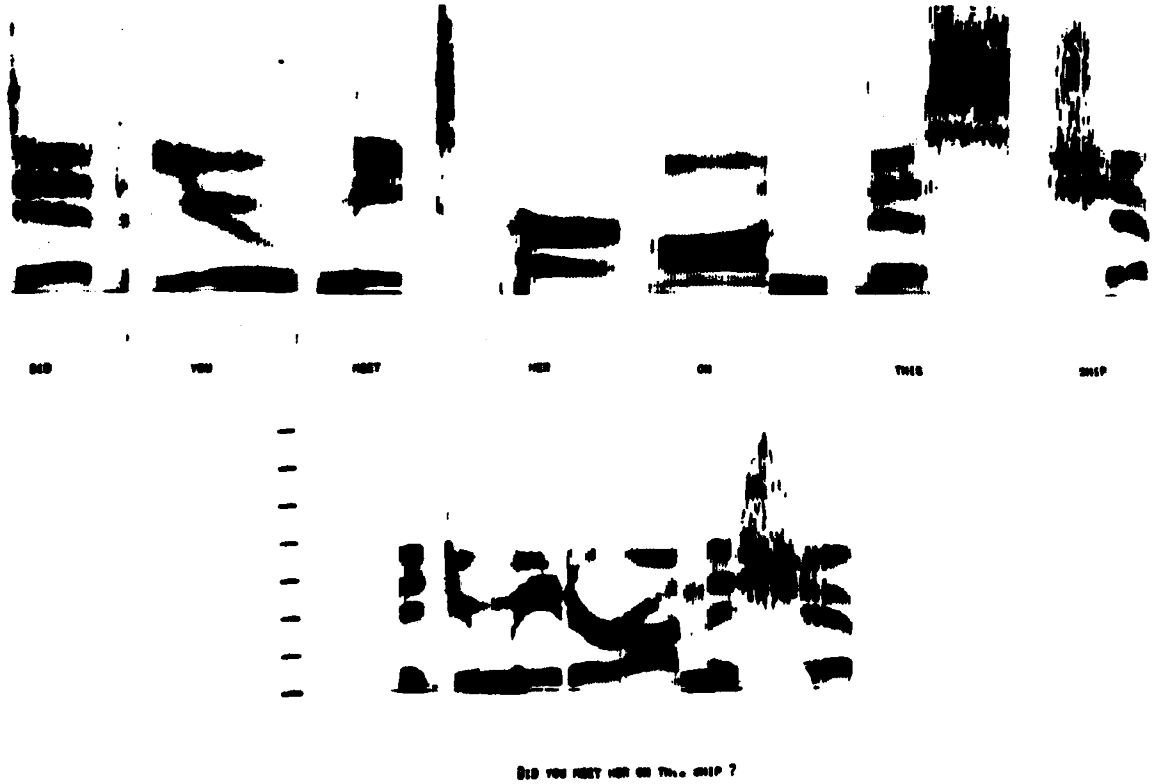


Fig. 2. Spectrogram of the words 'did', 'you', 'meet', 'her', 'on', 'this', 'ship', spoken in isolation and in a sentence by a male speaker. The spectrogram illustrates phonological processes such as palatalization and flapping that can operate at word boundaries.

drawbacks with such a description. The rules assume that the output units are discrete, whereas the variations seem to take on a continuous range of values. In addition, the application of context-dependent rules often requires the specification of the correct context, a process that can be prone to error. (For example, the identification of a retroflexed /t/ in the word 'tree' depends upon correctly identifying the following retroflexed consonant /r/). Finally there is no definitive agreement among researchers regarding the constituent units used to describe such variations, be it segment, syllable, or metrical foot.

The third type of phonetic rules specifies the optional realizations of a particular phonetic string. Most of the low-level *phonological* rules that describe alternate pronunciations of words fall into this category. These rules specify, for example, that a word such as 'international' can have many pronunciations including the deletion of /t/ and/or the deletion of the penultimate schwa. Traditionally, this problem is solved by expanding the lexicon, and the possible word combinations, with phonological rules to include all possible pronunciations (Cohen and Mercer, 1975; Woods and Zue, 1976). This approach also has some drawbacks. For example, dictionary expansion does not capture the nature of phonetic variability, namely that certain segments of a word are highly variable while others are relatively invariant. It is also difficult to assign a likelihood measure to each of the pronunciations. Finally, storing all alternate pronunciations is computationally expensive, since the size of the lexicon can increase substantially.

It is interesting to note that, for American English at least, most of the phonological rules tend to apply to unstressed syllables. Since the acoustic cues for phonetic segments around unstressed syllables are usually far less reliable than around stressed syllables [see, for example, Cutler and Foss (1977)], one may ask whether detailed knowledge of the various pronunciations are necessary for speech recognition. A recent study conducted by Huttenlocher and Zue (1983) indicates that phonetic segments within unstressed syllables provide little constraint for lexical access. Like in Shipman and Zue (1982), the phonemes were mapped into broad phonetic classes, except this time the mapping was done only for phonemes within stressed syllables. The entire unstressed syllable was mapped into a 'place holder' symbol, [*]. Thus, for example, the word 'spectrogram'/spɛktrogræm/ is represented by the pattern:

[strong fricative][stop][vowel][stop][*]
[stop][liquid or glide][vowel][nasal].

It was found that such representation still provided powerful constraints for lexical access. These results suggest that low-level phonological variations may be handled by 'wild-carding' the unstressed syllables where the functional load carried by the phonetic segments may be minimal.

4. Summary

In our view the speech signal is the output of a highly constrained production mechanism. The decoding, or recognition, of sentences involves the proper utilization of constraints at various levels, including acoustic-phonetic, phonological, lexical, syntactic, and semantic. In this paper, we discussed the types of constraints that exist at the phonetic and phonological levels, and demonstrated that such constraint must be captured in an automatic speech recognition system.

Over the past two decades, we have made significant improvements in our *qualitative* understanding of the phonetic and phonological constraints. However, there still remains a great deal of work that needs to be done. First, we need to study a sufficient amount of data so that these phenomena can be quantified. Second, a formal mechanism for describing these constraints, both in terms of the proper units and the proper grammar, must be devised. Finally, the interaction of these constraints at different levels must be proposed, tested, and incorporated. From a functional standpoint, there exist a variety of ways to handle the phonetic variability, including the use of probabilistic modeling. However, if we view the speech recognition problem as one of constructing a computational model for speech perception, then the identification, quantification, and formulation of these phonetic rules is a task that the research community must collectively undertake. It is only with such a long range research effort that we can, one day, hope to construct speech recognition systems with capabilities approaching that of humans.

Acknowledgement

Research supported by the Office of Naval Research under contract N00014-82-K-0727 and by the System Development Foundation.

References

S.E. Blumstein and K.N. Stevens (1979), "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," *J. Acoust. Soc. Am.*, Vol. 66, No. 4, pp. 1001–1017.

K.W. Church (1983), "Phrase-structure parsing: A method for taking advantage of allophonic constraints," Ph. D. thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

P.S. Cohen and R.L. Mercer (1975), "The phonological component of an automatic speech recognition system," in: D.R. Reddy, ed., *Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium*, Academic Press, New York, pp. 275–320.

R.A. Cole and V.W. Zue (1980), "Speech as eyes see it," Chapter 24 in: R.S. Nickerson, ed., *Attention and Performance VIII*, Lawrence Erlbaum Asso., Hillsdale, NJ pp. 475–494.

R.A. Cole, A.I. Rudnicky, V.W. Zue and D.R. Reddy (1980), "Speech as patterns on paper," Chapter 1 in: R.A. Cole, ed., *Perception and Production of Fluent Speech*, Lawrence Erlbaum Asso., Hillsdale, NJ, pp. 3–50.

A. Cutler and D.J. Foss (1977), "On the role of sentence stress in sentence processing," *Language and Speech*, Vol. 20, pp. 1–10

D.P. Huttenlocher and V.W. Zue (1983), "Phonetic and lexical constraints in speech recognition," paper presented at the *1983 AAAI International Conf.*, Washington, D.C.

I. Kameny (1975), "Comparison of formant spaces of retroflexed and nonretroflexed vowels," *IEEE Trans. Acoust. Speech Signal Process.* Vol. ASSP-23, pp. 38–49.

N.S.-Y. Kiang (1980), "Processing of speech by the auditory nervous system," *J. Acoust. Soc. Am.*, Vol. 68, pp. 830–835.

D.H. Klatt (1975), "Voice onset time, frication and aspiration in word-initial consonant clusters," *J. Speech Hearing Research*, Vol. 18, pp. 686–706.

L. Nakatani and K. Dukes (1977), "Locus of segmental cues for word juncture," *J. Acoust. Soc. Am.*, Vol. 62, No. 3, pp. 714–719.

B.T. Oshika, V.W. Zue, R.V. Weeks, H. Nue and J. Aurbach (1975), "The role of phonological rules in speech understanding research," *IEEE Trans. Acoust. Speech, signal Process.*, Vol. ASSP-23, pp. 104–112.

D.W. Shipman and V.W. Zue (1982), "Properties of large lexicons: Implications for advanced isolated word recognition systems," Conference Record, *1982 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 546–549.

N. Umeda (1975), "Vowel duration in American English," *J. Acoust. Soc. Am.*, Vol. 58, pp. 434–445.

N. Umeda (1977), "Consonant duration in American English," *J.* Acoust. Soc. Am., Vol. 61, pp. 846–858.

W. Woods and V.W. Zue (1976), "Dictionary expansion via phonological rules for a speech understanding system," Conference Record, *1975 IEEE Int. Conf. Acoust. Speech Signal Process.*, Philadelphia, PA, pp. 561–564.

V.W. Zue (1976), "Acoustic characteristics of stop consonants: A controlled study," Sc. D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology; also published by the University of Indiana Linguistic Club.

V.W. Zue and R.A. Cole (1979), "Experiments on spectrogram reading," Conference Record, 1979, *1979 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 116–119.

V.W. Zue and M. Laferriere (1979), "Acoustic study of medial /t, d/ in American English," *J. Acoust. Soc. Am.*, Vol. 66, No. 4, pp. 1039–1050.

V.W. Zue and S. Shattuck-Hufnagel (1980), "Palatalization of /s/ in American English: When is a /š/ not a /š/?", *J. Acoust. Soc. Am.*, Vol. 67, p. S27.

PROPOSAL FOR AN ISOLATED-WORD RECOGNITION SYSTEM BASED ON PHONETIC KNOWLEDGE AND STRUCTURAL CONSTRAINTS

Victor W. Zue

Room 36-549
Massachusetts Institute of Technology
Cambridge, MA 02139 U.S.A.

Abstract for an invited paper for a special session on "Human and Automatic Speech Recognition" at the Tenth International Congress of Phonetic Science, 1-6 August, 1983, Utrecht, the Netherlands.

Research supported by the Office of Naval Research under contract N00014-82-K-0727 and by the Sustem Development Foundation.

During the past decade, significant advances have been made in the field of isolated word recognition (IWR). In many instances, transitions from research results to practical implementations have taken place. Today, speech recognition systems that can recognize a small set of isolated words, say 50, for a given speaker with an error rate of less than 5% appear to be relatively common. Most of the current systems utilize little or no speech-specific knowledge, but derive their power from general-purpose pattern recognition techniques. The success of these systems can at least in part be attributed to the introduction of novel parametric representations (Makhoul, 1975), distance metrics (Itakura, 1975), and the very powerful time alignment procedure of dynamic programming (Sakoe and Chiba, 1971).

While we have clearly made significant advances in dealing with a small portion of the speech recognition problem, there is serious doubt regarding the extendibility of the pattern matching approach to tasks involving multiple speakers, large vocabularies and/or continuous speech. One of the limitations of the template matching approach is that both computation and storage grow (essentially) linearly with the size of the vocabulary. When the size of the vocabulary is very large, e.g., over 10,000 words, the computation and storage requirements associated with current IWR systems become prohibitively expensive. Even if the computational cost were not an issue, the performance of these IWR systems for a large vocabulary would surely deteriorate (Keilin et al., 1981). Furthermore, as the size of the vocabulary grows, it becomes imperative that such systems be able to operate in a speaker-independent mode, since training of the system for each user will take too long.

This paper proposes a new approach to large-vocabulary, isolated word recognition which combines detailed acoustic-phonetic knowledge with constraints on the sound patterns imposed by the language. The proposed system draws on the results of two sets of studies; one demonstrating the richness of phonetic information in the acoustic signal and the other demonstrating the power of structural constraints imposed by the language.

Spectrogram Reading Reliance on general pattern matching techniques has been partly motivated by the unsatisfactory performance of early phonetically-based speech recognition systems. The difficulty of automatic acoustic-phonetic analysis has also led to the speculation that phonetic information must be derived, in large part, from semantic, syntactic and discourse constraints rather than from the acoustic signal. For the most part, the poor performance of these phonetically-based systems can be attributed to the fact that our knowledge of the context-dependency of the acoustic characteristics of speech sounds was very limited at the time. However, this picture is slowly changing. We now have a far better understanding of contextual influences on phonetic segments. This improved understanding has been demonstrated in a series of spectrogram reading experiments (Cole et al. 1980). It was found that a trained subject can phonetically transcribe unknown sentences from speech spectrograms with an accuracy of approximately 85%. This performance is better than the phonetic recognizers reported in the literature, both in accuracy and rank order statistics. It was also demonstrated that the process of spectrogram reading makes use of explicit acoustic phonetic rules, and that this skill can be learned by others. These results suggest that the acoustic signal is rich in phonetic information, which should permit substantially better performance in automatic phonetic recognition.

However, even with a substantially improved knowledge base, a completely bottom-up phonetic analysis still has serious drawbacks. It is often difficult to make fine phonetic distinctions (for example, distinguishing the word pair "Sue/shoe") reliably across a wide range of speakers. Furthermore, the application of context-dependent rules often requires the specification of the correct context, a process that can be prone to error. (For example, the identification of a retroflexed /t/ in the word "tree" depends upon correctly identifying the retroflex consonant /r/.) Problems such as these suggest that a **detailed** phonetic transcription of an unknown utterance may not by itself be a desirable aim for the early application of phonetic knowledge.

Constraints on Sound Patterns Detailed segmental representation of the speech signal constitutes but one of the sources of encoded phonetic information. The sound patterns of a given language are not only limited by the inventory of basic sound units, but also by the allowable combinations of these sound units. Knowledge about such phonotactic constraints is presumably very useful in speech communication, since it provides native speakers with the ability to fill in phonetic details that are otherwise not available or are distorted. Thus, as an extreme example, a word such as "splint" can be recognized without having to specify the detailed acoustic characteristics of the phonemes /s/, /p/, and /n/. In fact, "splint" is the only word in the Merriam Pocket Dictionary (containing about 20,000 words) that satisfies the following description:

[CONS] [CONS] [1] [VOWEL] [NASAL] [STOP].

In a study of the properties of large lexicons, Shipman and Zue (1982) found that knowledge of even broad specification of the sound patterns of American English words, both at the segmental and suprasegmental levels, imposes strong constraints on their phonetic identities. For example, if each word in the lexicon is represented only in terms of 6 broad manner categories (such as vowel, stop, strong fricative, etc.), then the average number of words in a 20,000-word lexicon that share the same sound pattern is about 2. In fact, such crude classification will enable about 1/3 of the lexical items to be uniquely determined.

There is indirect evidence that the broad phonetic characteristics of speech sounds and their structural constraints are utilized to aid human speech perception. For example, Blesser (1969) has shown that people can be taught to perceive spectrally-rotated speech, in which manner cues and suprasegmental cues are preserved while detailed place cues are severely distorted. The data on misperception of fluent speech reported by Bond and Garnes (1980) and the results of experiments on listening for mispronunciation reported by Cole and Jakimik (1980) also suggest that the perceptual mechanism utilizes information about the broad phonetic categories of speech sounds and the constraints on how they can be combined.

Proposed System Based on the results of the two studies cited above, we propose a new approach to phonetically-based isolated-word recognition. This approach is distinctly different from previous attempts in that detailed phonetic analysis of the acoustic signal is not performed. Rather, the speech signal is segmented and classified into several broad manner categories. The broad phonetic (manner) classifier serves several purposes. First, errors in phonetic labeling, which are most often caused by detailed phonetic analyses, would be reduced. Second, by avoiding fine phonetic distinctions, the system should also be less sensitive to interspeaker variations. Finally, we speculate that the sequential constraints and their distributions, even at the broad phonetic level, may provide powerful mechanisms to reduce the search space substantially. This last feature is particularly important when the size of the vocabulary is large (of the order of several thousand words or more).

Once the acoustic signal has been reduced to a string (or lattice) of phonetic segments that have been broadly classified, the resulting representation will be used for lexical access. The intent is to reduce the number of possible word candidates by utilizing knowledge about the structural constraints, both segmental and suprasegmental, of the words. The result, as indicated previously, should be a relatively small set of word candidates. The correct word will then be selected through judicious applications of detailed phonetic knowledge.

In summary, this paper presents a new approach to the problem of recognizing isolated words from large vocabularies and multiple speakers. The system initially classifies the acoustic signal into several broad manner categories. Once the potential word candidates have been significantly reduced through the utilization of the structural constraints, then a detailed examination of the acoustic differences would follow. Such a procedure will enable us to deal with the large vocabulary recognition problem in an efficient manner. What is even more important is the fact that such an approach bypasses the often tedious and error-prone process of deriving a complete phonetic transcription from the acoustic signal. In this approach, detailed acoustic phonetic knowledge can be applied in a top-down verification mode, where the exact phonetic context can be specified.

REFERENCES

Bond, Z.S. and Garnes S. (1980) "Misperceptions of Fluent Speech," Chapter 5 in **Perception and Production of Fluent Speech**, ed. R.A. Cole, 1154-132 (Lawrence Erlbaum Asso., Hillsdale, New Jersey).

Cole, R.A. and Jakimik, J. (1980) "A Model of Speech Perception," Chapter 6 in **Perception and Production of Fluent Speech**, ed. R.A. Cole, 133-163 (Lawrence Erlbaum Asso., Hillsdale, New Jersey).

Cole, R.A., Rudnicky, A.I., Zue, V.W., and Reddy, D.R. (1980) "Speech as Patterns on Paper," Chapter 1 in **Perception and Production of Fluent Speech**, ed. R.A. Cole, 3-50 (Lawrence Erlbaum Asso., Hillsdale, New Jersey).

Itakura, F. (1975) "Minimum Prediction Residual Principle Applied to Speech Recognition," **IEEE Trans. Acoustics, Speech, and Signal Processing**, Vol. ASSP-23, 67-72.

Keilin, W.J., Rabiner, L.R., Rosenberg, A.E., and Wilpon, J.G. (1981) "Speaker Trained Isolated Word Recognition on a Large Vocabulary," **J. Acoust. Soc. Am.**, Vol. 70, S60.

Makhoul, J.I. (1975) "Linear Prediction: A Tutorial Review," **Proc. IEEE**, Vol. 63, 561-580.

Shipman, D.W. and Zue, V.W. (1982) "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," **Conference Record, IEEE 1982 International Conference on Acoustics, Speech and Signal Processing**, 546-549.

Sakoe, H. and Chiba, S. (1971) "A Dynamic Programming Optimization for Spoken Word Recognition," **IEEE Trans. Acoustics, Speech, and Signal Processing**, Vol. ASSP-26, 43-49.

Presented at the American Association For Artificial Intelligenc
Annual Conference, August 21-26, 1983

PHONOTACTIC AND LEXICAL CONSTRAINTS IN SPEECH RECOGNITION

Daniel P. Huttenlocher and Victor W. Zue
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

We demonstrate a method for partitioning a large lexicon into small equivalence classes, based on sequential phonetic and prosodic constraints. The representation is attractive for speech recognition systems because it allows all but a small number of word candidates to be excluded, using only gross phonetic and prosodic information. The approach is a robust one in that the representation is relatively insensitive to phonetic variability and recognition error.

INTRODUCTION

Speech is the output of a highly constrained system. While it has long been recognized that there are multiple sources of constraint on speech production and recognition, natural language research has tended to focus on the syntactic, semantic, and discourse levels of processing. We believe that constraints at the phonological and lexical levels, although less well understood, are as important in recognition as higher level constraints. For a given language, the speech signal is produced with a limited inventory of possible sounds, and these sounds can only be combined in certain ways to form meaningful words. Knowledge about such constraints is implicitly possessed by native speakers of a given language. For example, an English speaker knows that "vnuk" is not an English word because it violates the *phonotactic* rules governing the allowable sound sequences of the language. He or she also knows that if an English word starts with three consonants, then the first consonant must be an /s/, and the second consonant must be either /p/, /t/, or /k/. On the other hand "smeck" is a permissible sequence of sounds in English, but is not a word because it is not in the lexicon. Such phonotactic and lexical knowledge is presumably important in speech recognition, particularly when the acoustic cues to a speech sound are missing or distorted. Perceptual data demonstrate the importance of these lower level phonological and lexical constraints. First, people are good at recognizing isolated words, where there are no higher-level syntactic or semantic constraints [3]. Second, trained phoneticians are rather poor at phonetically transcribing speech from an unknown language, for which they do not possess the phonotactic and lexical knowledge [11].

Perceptual data demonstrate that phonotactic and lexical knowledge is useful in speech recognition. we are concerned with *how* such knowledge can be used to constrain the recognition task. In this paper we investigate some phonotactic and lexical constraints by examining certain properties of large lexicons. First we consider the effects of representing words in terms of broad phonetic classes rather than specific phones. Then we discuss how this representation handles some common problems in speech recognition such as acoustic variability, and segment deletion.

PHONOTACTIC CONSTRAINTS CAN BE EXTREMELY USEFUL IN LEXICAL ACCESS

Most of the phonological rules informally gathered by linguists and speech researchers are specified in terms of broad phonetic classes rather than specific phones. For example, the homorganic rule of nasal-stop clusters specifies that nasals and stop consonants must be produced at the same place of articulation. Thus we have words like "limp" or "can't", but not "limt" or "camp". In speech perception, there is also evidence that people use knowledge about the broad classifications of speech sounds. For example, the non-word "shpeech" is still recognizable as the word "speech", while "tpeech" is not. This is because "s" and "sh" both belong to the same class of sounds (the strong fricatives), while "t" belongs to a different class (the aspirated stops). The perceptual similarity of these broad phonetic classes has long been known [9]. These broad classes are based on the so called *manner* of articulation differences. For example, the stop consonants /p/, /t/, and /k/ are all produced in the same manner, with closure, release and aspiration. The stops differ from one another in their respective *place of articulation*, or the shape of the vocal tract and position of the articulators. Manner differences tend to have more robust and speaker-invariant acoustic cues than place differences [8]. This makes broad manner classes attractive for recognition systems. However, until quite recently little was known about the role these constraints play in recognition [1] [14]. Therefore, speech recognition and understanding systems have not made much use of this information [4] [5].

Although the importance of phonotactic constraints has long been known, the magnitude of their predictive power was not apparent until Shipman and Zue reported a set of studies recently [10]. These studies examined the phonotactic constraints of American English from the phonetic distributions in the 20,000-word Merriam Webster's Pocket Dictionary. In one

* Research supported by the Office of Naval Research under contract N00014-82-K-0727 and by the System Development Foundation

study the phones of each word were mapped into one of six broad phonetic categories: vowels, stops, nasals, liquids and glides, strong fricatives, and weak fricatives. Thus, for example, the word "speak", with a phonetic string given by /spik/, is represented as the pattern:

[strong-fricative][stop][vowel][stop]

It was found that, even at this broad phonetic level, approximately 1/3 of the words in the 20,000-word lexicon can be uniquely specified. One can view the broad phonetic classifications as partitioning the lexicon into equivalence classes of words sharing the same phonetic class pattern. For example, the words "speak" and "steep" are in the same equivalence class. The average size of these equivalence classes for the 20,000-word lexicon was found to be approximately 2, and the maximum size was approximately 200. In other words, in the worst case, a broad phonetic representation of the words in a large lexicon reduces the number of possible word candidates to about 1% of the lexicon. Furthermore, over half of the lexical items belong to equivalence classes of size 5 or less. This distribution was found to be fairly stable for lexicons of about 2,000 or more words; for smaller lexicons the specific choice of words can make a large difference in the distribution.

HOW ROBUST IS A BROAD PHONETIC REPRESENTATION?

The above results demonstrate that broad phonetic classifications of words can, in principle, reduce the number of word candidates significantly. However, the acoustic realization of a phone can be highly variable, and this variability introduces a good deal of recognition ambiguity in the initial classification of the speech signal [6] [7] [12]. At one extreme, the acoustic characteristics of a phoneme can undergo simple modifications as a consequence of contextual and inter-speaker differences. Figure 1 illustrates the differences in the acoustic signal for the various allophones of /t/ in the words "tree", "tea", "city", and "beaten". At the other extreme, contextual effects can also produce severe modifications in which phonemes or syllables are deleted altogether. Thus, for example, the word "international" can have many different realizations, some of which are illustrated in Figure 2. Not only may phonemes be deleted, some pronunciations of a word may even have a different number of syllables than the clearly enunciated version.

In order to evaluate the viability of a broad phonetic class representation for speech recognition systems, two major problems must first be considered. The first problem is that of mis-labeling a phonetic segment, and the second problem is the deletion of a segment altogether. It is important to note that these phenomena can occur as a consequence of the high level of variability in natural speech, as well as resulting from an error by the speech recognition system. That is, not only can the recognizer make a mistake, a given speaker can utter a word with changed or deleted segments. Therefore, even a perfect recognizer would still have "errors" in its input. We address segmental variation and segmental deletion errors in the next two sections.

The scheme proposed by Shipman and Zue can handle allophonic variations, such as the different realizations of /t/.

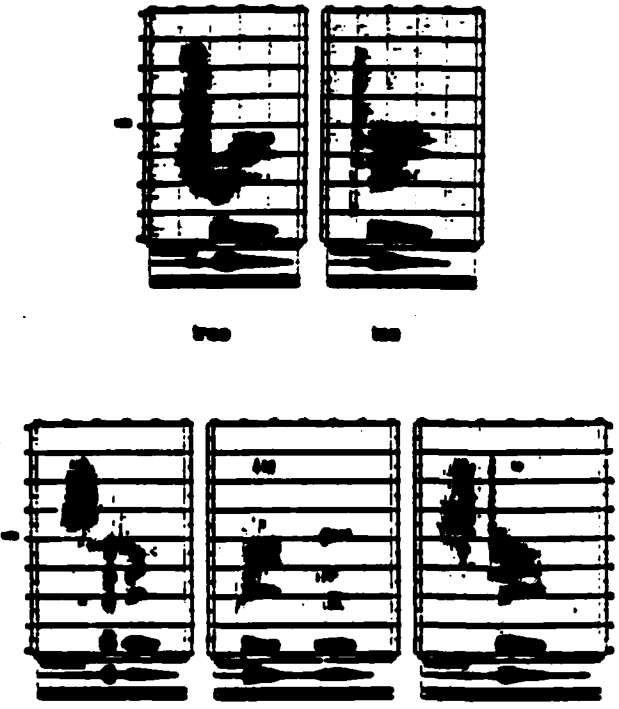


Figure 1:
Spectrograms Illustrating the Acoustic Realizations of the Various Allophones of /t/

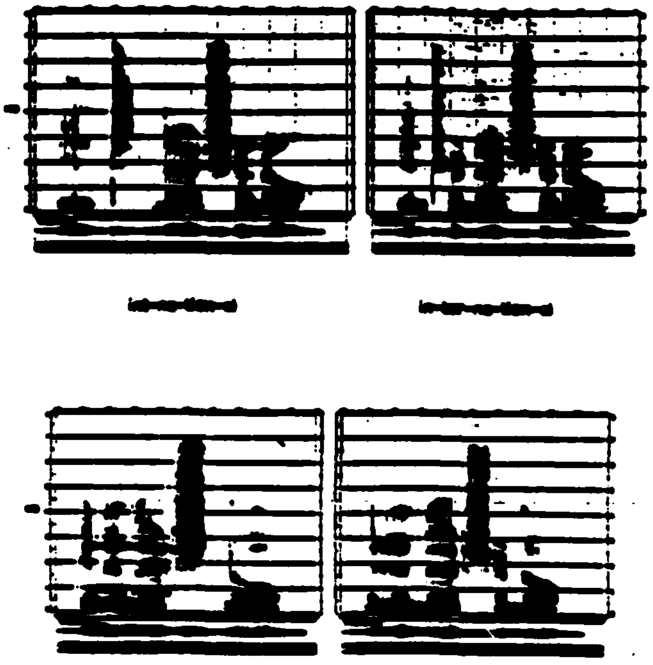


Figure 2:
Spectrograms Illustrating Several Possible Pronunciations for the Word "International"

This is because contextual variations tend to affect the detailed acoustic realizations of the phonetic segments, as opposed to the gross manner features used in the broad classes. When accessing the lexicon based on broad phonetic classification, detailed allophonic differences are completely disregarded. On the other hand, some uncertainties due to inter-speaker differences and recognizer errors are bound to occur. Given such uncertainties, one may ask whether the original results of Shipman and Zue still hold for lexical access.

There are a number of ways such a question can be answered. In one study we inferred the effect of these labeling ambiguities by allowing a fixed percentage of the phonetic segments in the lexicon to be unclassified while assuming that the remaining segments are classified correctly. Thus, for example, a 10% phonetic uncertainty is simulated by assuming that 90% percent of the phonetic segments in the lexicon are classified correctly. The remaining 10% of the segments can effectively be matched to any of the six phonetic categories. In order to accommodate such ambiguities, words in the lexicon must now contain not only the correct broad phonetic representation, but also those representations resulting from including unclassified segments. Admittedly our assumptions are not completely realistic, since labeling uncertainties do not occur for only a fixed percentage of the segments. Furthermore, labeling uncertainties usually arise among subsets of the broad categories. For example, it may be possible to confuse a strong fricative with a weak one, but not a strong fricative with a vowel. Nevertheless, we believe that such a simulation provides a glimpse of the effect of labeling uncertainties.

Table 1 compares the lexical distributions obtained from the original results of Shipman and Zue (in the first column) with those obtained by allowing 10% and 20% labeling uncertainty (in the second and third columns). The results indicate that, even allowing for a good deal of classification ambiguity, lexical constraints imposed by sequences of broad phonetic classes are still extremely powerful. In all cases, over 30% of the lexical items can be uniquely specified, and over 50% of the time the size of the equivalence class is 5 or less. On the other hand, the maximum sizes of the equivalence classes grow steadily as the amount of labeling uncertainty increases.

	Whole Word	10% Label. Errors	20% Label. Errors
% Uniquely Specified	32%	32%	32%
% In Classes of Size 5 or Less	56%	56%	56%
Max Class Size	210	278	348

Table 1: Comparison of Lexical Constraint with and without Labeling Uncertainty

PROSODIC INFORMATION CAN ALSO AID LEXICAL ACCESS

The broad phonetic class representation cannot handle segment or syllable deletions, since when a segment deletion occurs, the broad phonetic class sequence is affected. Traditionally, this problem is solved by expanding the lexicon via phonological rules, in order to include all possible pronunciations of each word [13]. We find this alternative unattractive for several reasons. For example, dictionary expansion does not capture the nature of phonetic variability. Once a given word is represented as a set of alternate pronunciations, the fact that certain segments of a word are highly variable while others are relatively invariant is completely lost. In fact, below we see that the less variable segments of a word provide more lexical constraint than those segments which are highly variable. Another problem with lexical expansion is that of assigning likelihood measures to each pronunciation. Finally, storing all alternate pronunciations is computationally expensive, since the size of the lexicon can increase substantially.

Some segments of a word are highly variable, while others are more or less invariant. Depending on the extent to which the variable segments constrain lexical access, it might be possible to represent words only in terms of their less variable parts. For instance, in American English most of the phonological rules apply to unstressed syllables. In other words, phonetic segments around unstressed syllables are more variable than those around stressed syllables. Perceptual results have also shown that the acoustic cues for phonetic segments around unstressed syllables are usually far less reliable than around stressed syllables [2]. Thus, one may ask to what extent phones in unstressed syllables are necessary for speech recognition.

In an attempt to answer this question, we compared the relative lexical constraint of phones in stressed versus unstressed syllables. In one experiment, we classified the words in the 20,000-word *Webster's Pocket Dictionary* either according to only the phones in stressed syllables, or according to only the phones in unstressed syllables. In the first condition, the phones in stressed syllables were mapped into their corresponding phonetic classes while the entire unstressed syllables were mapped into a "placeholder" symbol. In the second condition the opposite was done. For example, in the first condition the word "paper", with the phonemic string /'pe'-pə/, is represented by the pattern:

[stop][vowel][*]

where * is the unstressed syllable marker. In the second condition the word is represented by the pattern:

[*][stop][vowel]

where * is the stressed syllable marker. Note that at first glance a significant amount of information is lost by mapping an entire syllable into a placeholder symbol. Closer examination reveals, however, that the placeholder symbol retains the prosodic structure of the words. A representation which makes this more explicit combines the partial phonetic classification with syllabic stress information. Thus, in the first condition, the word "paper" would be represented as:

[stop][vowel] + [S][U]

where [S] and [U] correspond to stressed and unstressed syllables, respectively.

The results of this experiment are given in the second two columns of Table 2. The results summarized in the table are obtained by explicitly representing the prosodic information as sequences of stressed and unstressed syllables. The results for "wildcarding" the deleted syllables are almost identical and hence are not presented here. The first column of the table gives the results for the whole word (as in [10]). The second and third columns show the cases where phonetic information is only preserved in the stressed or in the unstressed syllables. It should be noted that the results cannot be accounted for simply on the basis of the number of phones in stressed versus unstressed syllables. For the entire lexicon, there are only approximately 1.5 times as many phones in stressed than in unstressed syllables. In addition, if one considers only polysyllabic words, there are almost equal numbers of phones in stressed and unstressed syllables, yet the lexical distribution remains similar to that in Table 2.

These results demonstrate that the phonotactic information in stressed syllables provides much more lexical constraint than that in unstressed syllables. This is particularly interesting in light of the fact that the phones in stressed syllables are much less variable than those in unstressed syllables. Therefore, recognition systems should not be terribly concerned with correctly identifying the phones in unstressed syllables. Not only is the signal highly variable in these segments, making classification difficult; the segments do not constrain recognition as much as the less variable segments.

This representation is very robust with respect to segmental and syllabic deletions. Most segment deletions, as was pointed out above, occur in unstressed syllables. Since the phones in unstressed syllables are not included in the representation, their deletion or modification is ignored. Syllabic deletions occur exclusively in unstressed syllables, and usually in syllables containing just a single phone. Thus, words with a single-phone unstressed syllable can be stored according to two syllabic stress patterns. For example the word "international" would be encoded by the phones in its stressed syllables:

[vowel][nasal][nasal][vowel][strong-fricative]

with the two stress patterns [S][U][S][U][U] and [S][U][S][U] for the 5- and 4-syllable versions. The common pronunciations of "international" (e.g., those in Figure 2) are all encoded by these two representations, while unreasonable pronunciations like "interashnel" are excluded.

SUMMARY

We have demonstrated a method for encoding the words in a large lexicon according to broad phonetic characterizations. This scheme takes advantage of the fact that even at a broad level of description, the sequential constraints on allowable sound sequences are very strong. It also makes use of the fact that the phonetically variable parts of words provide much less lexical constraint than the phonetically invariant parts. The interesting properties of the representation are that it is based on relatively robust phonetic classes, it allows for phonetic variability, and it partitions the lexicon into very small equivalence classes. This makes the representation attractive for speech recognition systems [15].

	Whole Word	Stressed Syls.	Un-Stressed Syls.
% Uniquely Specified	32%	17%	8%
% In Classes of Size 5 or Less	56%	38%	19%
Average Class Size	2.3	3.8	7.7
Max Class Size	210	201	3717

Table 2: Comparison of Lexical Constraint in Stressed vs Unstressed Syllables

Using a broad phonetic representation of the lexicon is a search avoidance technique, allowing a large lexicon to be pruned to a small set of potential word candidates. An essential property of such a technique is that it retains the correct answer in the small candidate set. We have demonstrated that, for a wide variety of speech phenomena, a broad phonetic representation has this property.

REFERENCES

[1] Broad, D.J. and Shoup, J.E. (1975) "Concepts for Acoustic Phonetic Recognition" in R.D. Reddy, *Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium*, Academic Press, New York.

[2] Cutler, A. and Foss, D.J. (1977) "On the Role of Sentence Stress in Sentence Processing", Language and Speech, Vol. 20, 1-10.

[3] Dreher, J.J. and O'Neill, J.J. (1957) "Effects of Ambient Noise on Speaker Intelligibility for Words and Phrases", Journal of the Acoustical Society of America, vol. 29, no. 12.

[4] Erman, L.D, Hayes-Roth, F., Lesser, V.R., and Reddy, R.D. (1980). "The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty", Computing Surveys, vol. 12, no. 2, 213-253.

[5] Klatt, D.H. (1977) "Review of the ARPA Speech Understanding Project", Journal of the Acoustical Society of America, vol. 62, no. 6, 1345-1366.

[6] Klatt, D.H. (1980) "Speech Perception: A Model of Acoustic Phonetic Analysis and Lexical Access" in R. Cole, *Perception and Production of Fluent Speech*. Lawrence Erlbaum Assoc., Hillsdale, N.J.

[7] Klatt, D.H. (1983) "The Problem of Variability in Speech Recognition and in Models of Perception". Invited paper at the 10th International Congress of Phonetic Sciences.

[8] Lea, W.A. (1980) *Trends in Speech Recognition*. Prentice-Hall, N.Y.

[9] Miller, G.A. and Nicely, P.E. (1954) "An Analysis of Perceptual Confusions Among Some English Consonants", Journal of the Acoustical Society of America, vol. 27, no. 2, 338-352.

[10] Shipman, D.W. and Zue, V.W. (1982) "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems", Conference Record, IEEE International Conference on Speech Acoustics and Signal Processing, Paris, France, 546-549.

[11] Shockey, L. and Reddy, R.D. (1974) "Quantitative Analysis of Speech Perception: Results from Transcription of Connected Speech from Unfamiliar Languages" *Speech Communication Seminar*, G. Fant (Ed).

[12] Smith, A. (1977) "Word Hypothesization for Large-Vocabulary Speech Understanding Systems", Doctoral Dissertation, Carnegie-Mellon University, Department of Computer Science.

[13] Woods, W. and Zue, V.W. (1976) "Dictionary Expansion via Phonological Rules for a Speech Understanding System", Conference Record, IEEE International Conference on Speech Acoustics and Signal Processing, Phila, Pa. 561-564.

[14] Zue, V.W. (1981) "Acoustic-Phonetic Knowledge Representation: Implications from Spectrogram Reading Experiments". Proceedings of the 1981 NATO Advanced Summer Institute on Speech Recognition, Bonas, France.

[15] Zue, V.W. and Huttenlocher, D.P. (1983) "Computer Recognition of Isolated Words from Large Vocabularies". IEEE Computer Society Trends and Applications Conference, Washington, D.C., 121-128.

Exploring Phonotactic and Lexical Constraints in Word Recognition [1]

by
Daniel P. Huttenlocher
and
Victor W. Zue

Room 36-541
Department of Electrical Engineering and Computer Science,
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

[1] Paper I2 presented at the 106th Meeting of the Acoustical Society of America, San Diego, CA, November 8, 1983

Exploring Phonotactic and Lexical Constraints in Word Recognition. Daniel P. Huttenlocher and Victor W. Zue (Room 36-549, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139)

In a previous meeting of the Society, Zue and Shipman demonstrated that the constraints imposed by the allowable sound sequences of a language are extremely powerful. Even at a broad phonetic level of representation, sequential constraints severely limit the number of possible word candidates [JASA Vol. 71, S7]. This provides an attractive model for lexical access based on partial phonetic information. However, Zue and Shipman's results did not take into account the fact that the acoustic realizations of phonetic segments are highly variable, and this variability introduces a good deal of recognition ambiguity in the initial classification of the signal. We have conducted a set of studies investigating the robustness of sequential phonetic constraints with respect to variability and error in broad phonetic classification. In these studies segment misclassifications or deletions are permitted. In one study it was found that the phonetically variable parts of words (around reduced syllables) provide much less lexical constraint than the phonetically invariant parts. Thus, by utilizing the information from robust parts of a word, a large lexicon can still be partitioned into small equivalence classes. Detailed results of the studies will be presented. [Work supported by the Office of Naval Research under contract N00014-82-K-0727 and by the System Development Foundation.]

Technical Committee: Speech Communication

Subject Classification Numbers: 43.70Sc, 43.70Ve

Telephone Number: (617) 253-8513 (V. Zue)

Send Acceptance or rejection notice to V. Zue

In a previous meeting of the Society, Zue and Shipman presented some results demonstrating the predictive power of phonotactic and lexical constraints in American English. Using the 20,000-word Merriam Webster·s Pocket Dictionary as their database, they mapped the phonemes of each word into one of six broad phonetic categories: vowels, stops, nasals, liquids and glides, strong fricatives, and weak fricatives. One can view the broad phonetic classifications as partitioning the lexicon into equivalence classes of words sharing the same phonetic class pattern. Thus, for example, the words "speak", "stop", "scout", and 48 other words fall into the same equivalence class: (STRONG-FRICATIVE) (STOP) (VOWEL) (STOP). Zue and Shipman found that the average size of these equivalence classes is approximately 2, and the maximum class size is approximately 200. In addition, it was found that, even at this broad phonetic level, approximately 1/3 of the words in the 20,000-word lexicon can be uniquely specified. One conclusion of their study is that lexical and phonotactic constraints are extremely powerful, even at the broad phonetic level.

Before presenting any new results, we would like to add two footnotes to the studies conducted by Zue and Shipman. First, the average equivalence class size may not have been an appropriate measure of lexical and phonotactic constraint. A more informative measure may be the expected value of the equivalence class. That is, given a word, what is the size of the equivalence class into which the word is likely to fall. We computed the expected value for the Zue and Shipman results and they are shown on the first overlay. The expected value, while greater by an order of magnitude than the average value, still represents only a tenth of a percent of the entire lexicon. As an indication of the spread of the distribution, we have also included the median class size.

It should be noted that these results are for a lexicon where the words are given uniform weighting. However, the frequency distribution of English words is highly skewed. When the words in the lexicon are weighted in terms of their frequency of occurrence based on the Brown Corpus, as shown on the second overlay, the median increases a good deal, whereas as the expected value only increases moderately. For the remainder of this presentation, the words in the lexicon have all been weighted by their frequency of occurrence in the Brown Corpus, thus providing a closer approximation to the usage of words in a language.

Our second comment on Zue and Shipman·s results is that they explored segmental/phonetic and prosodic constraints independently of one another. Since these sources of information are presumably useful in combination, we augmented the broad phonetic representation with two-level stress information for the word. Thus, for example, the word "piston" is represented both by a broad phonetic classification of: (STOP) (VOWEL) (STRONG-FRICATIVE) (STOP) (VOWEL) (NASAL), and a prosodic representation of: (STRESSED) (UNSTRESSED). The results of incorporating stress information are shown on the next viewgraph. As can be seen, the median, the expected value and the maximum equivalence class size all decreased as a result of introducing lexical stress information, suggesting the usefulness of prosodic information.

The above results demonstrate that broad phonetic classifications of words can, in principle, reduce the number of word candidates significantly. However, the acoustic realization of a phone can be highly variable, and this variability introduces a good deal of recognition ambiguity in the initial classification of the speech signal. At one extreme, the acoustic characteristics of a phoneme can undergo simple modifications as a consequence of contextual and inter-speaker differences. At the other extreme, contextual effects can also produce severe modifications in which phonemes or syllables are deleted altogether. Thus, for example, the word "international" can have many different realizations, including the deletion of the second as well as penultimate syllables.

In order to evaluate the viability of a broad phonetic class representation for speech recognition systems, two major problems must first be considered. The first problem is that of mis-labeling a phonetic segment, and the second problem is the deletion of a segment altogether. It is important to note that these phenomena can occur resulting from an error by the speech recognition system, and as a consequence of the high level of variability in natural speech. That is, not only can the recognizer make a mistake, a given speaker can utter a word with changed or deleted segments. Therefore, even a perfect recognizer would still have "errors" in its input. The primary focus of this talk is to investigate the effects on Zue and Shipman's results when errors, both in classification and segmentation, are introduced.

We have tried to infer the effect of mis-labeling phonetic segments, by allowing for reasonable confusions among the six phonetic classes. The allowable confusion is determined based on our acoustic-phonetic intuitions, as well as our past experience with speech recognition front-ends. Some of the confusions are context-independent, whereas others are permitted only under certain phonetic environments. In one study, we allowed strong fricatives to be confused with weak fricatives while only medial nasals can be confused with with liquids and glides, and only word initial aspirated stops can be confused with strong fricatives. Thus, in this study each word is represented by one or more broad phonetic sequences, depending on the possible confusions in the word. In the next viewgraph we see that introducing these confusions did not change the previous results significantly. This suggests that, if the classification uncertainty is reasonable, then lexical constraints imposed by sequences of broad phonetic classes are still extremely powerful.

We now turn to the second issue, namely, segmentation uncertainty due to front-end error or alternate pronunciations. The broad phonetic representation cannot handle segment or syllable deletions, because when a segment is deleted the broad phonetic class sequence is affected. Traditionally, this problem is solved by expanding the lexicon via phonological rules, in order to include all possible pronunciations of each word. We find this alternative unattractive for several reasons. First of all, dictionary expansion does not capture the nature of phonetic variability. Once a given word is represented as a set of alternate pronunciations, the fact that certain segments of a word are highly variable while others are relatively invariant is lost. In fact, we shall see that the less variable segments of a word provide more lexical constraint than those segments which are highly variable. Another problem with lexical expansion is that of assigning realistic likelihood measures to each pronunciation. Finally, storing all alternate pronunciations is computationally expensive, since the size of the lexicon can increase substantially.

Some segments of a word are highly variable, while others are more or less invariant. Depending on the extent to which the variable segments constrain lexical access, it might be possible to represent words only in terms of their less variable parts. It is interesting to note that, in American English, most of the low-level phonological rules apply to unstressed syllables. In other words, phonetic segments around unstressed syllables are more variable than those around stressed syllables. Perceptual results have also shown that the acoustic cues for phonetic segments around unstressed syllables are usually far less reliable than around stressed ones. Thus, one may ask to what extent the phones in unstressed syllables are useful for lexical access.

In an attempt to answer this question, we compared the relative lexical constraint of phones in stressed versus unstressed syllables. In one experiment, we classified the words in the 20,000-word Webster·s Pocket Dictionary either according to only the phones in stressed syllables, or according to only the phones in unstressed syllables. Lexical representations for this experiment are illustrated, for the word "piston", in the next viewgraph. In the first condition, shown on the left, the phones in stressed syllables were mapped into their corresponding phonetic classes while each unstressed syllable was mapped into a placeholder symbol. In the second condition, shown on the right, the opposite was done. For both conditions, stress information is retained.

The results of this experiment are given in the next viewgraph. By comparing the outcomes of the two conditions, it can be seen that information within stressed syllables provides much more constraints for lexical access than that in unstressed syllables. This is particularly interesting in light of the fact that the phones in stressed syllables are much less variable than those in unstressed syllables. Therefore, recognition systems may not have to be terribly concerned with correctly identifying the phones in unstressed syllables. Not only is the signal highly variable in these segments, making classification difficult; the segments do not constrain recognition as much as the less variable segments.

This representation is very robust with respect to segmental and syllabic deletions. As pointed out previously, most segment deletions occur in unstressed syllables. Since the phones in unstressed syllables are not included in the representation, their deletion or modification is ignored.

In summary, we have demonstrated a method for encoding the words in a large lexicon using broad phonetic and prosodic information. This scheme takes advantage of the fact that even at a broad level of description, the sequential constraints on allowable sound sequences are very strong. It also makes use of the fact that the phonetically variable parts of words provide much less lexical constraint than the phonetically invariant parts. The interesting properties of the representation are that it is based on relatively robust phonetic classes, it allows for phonetic variability, and it partitions the lexicon into very small equivalence classes. This makes the representation attractive as a search avoidance techniques for large-vocabulary speech recognition systems.

Representing Words By Broad Phonetic Classes

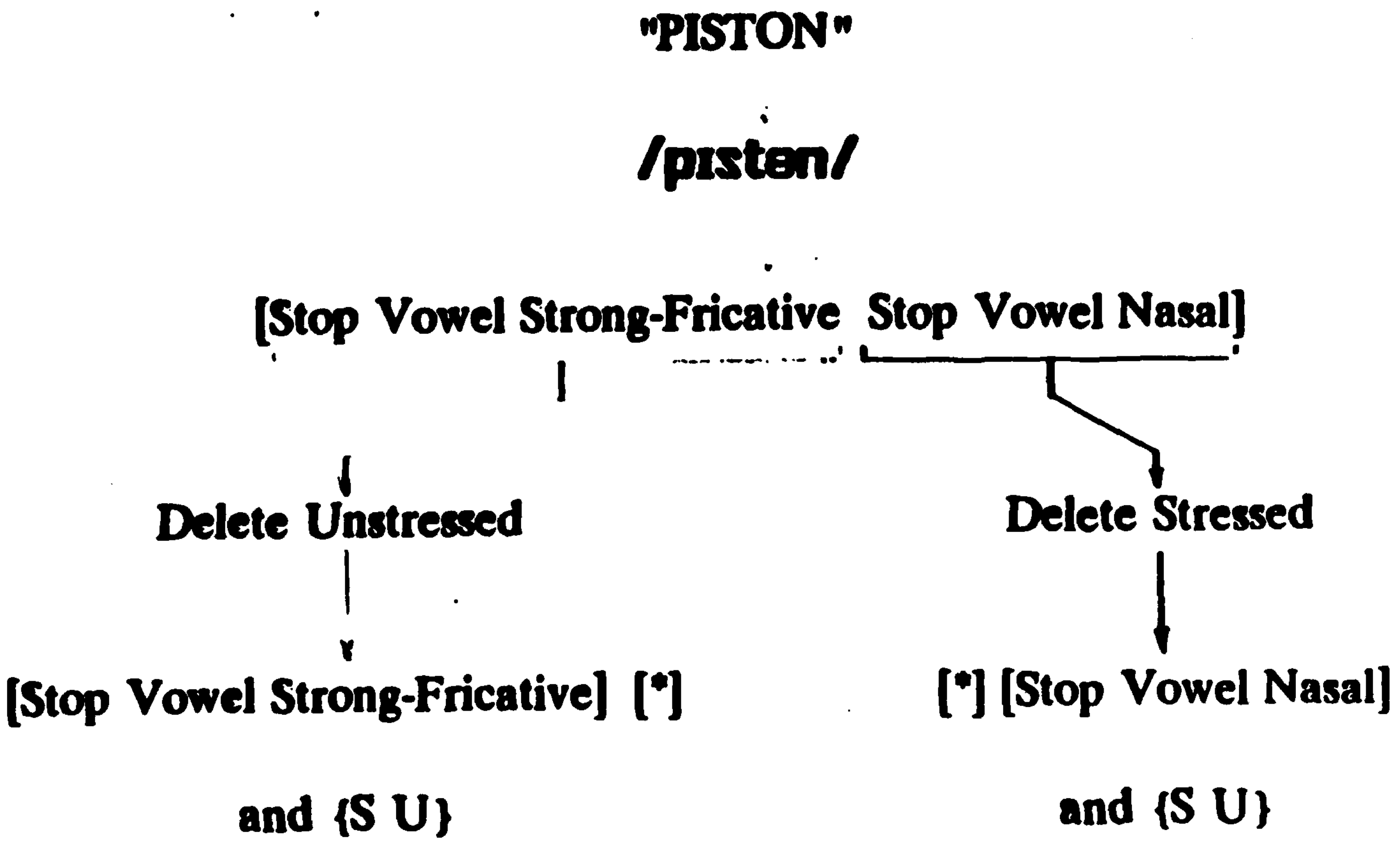
	Zue & Shipman	Z & S (Freq. Weighted)
Average Class Size	2	
Maximum Class Size	223	223
%Lexicon Unique	32	6
Expected Class Size	22	34
Median Class Size	4	25

Combining Segmental and Prosodic Constraints

	Segmental Only	Segmental + Stress
Expected Class Size	34	18
Median Class Size	25	2
Maximum Class Size	223	209
% Lexicon Unique	6	39

Allowing for Reasonable Labelling Confusions

	No Confusions	With Confusions
Expected Class Size	26	40
Median Class Size	8	9
Maximum Class Size	209	233
%Lexicon Unique	27	26



"PISTON"

/pɪstən/

[Stop Vowel Strong-Fricative Stop Vowel Nasal]

Delete Unstressed

[Stop Vowel Strong-Fricative] [*]

and {S U}

Delete Stressed

[*] [Stop Vowel Nasal]

and {S U}

Deletion Of Phonetic Information
Around Unstressed Or Stressed Syllables

	No Deletion	Delete Unstressed	Delete Stressed
Expected Class Size	26	40	2013
Median Class Size	8	22	1725
Maximum Class Size	223	261	3703

Exploring Allophonic and Lexical Constraints in a Continuous Speech Recognition System [1]

by
Francine R. Chen
and
Victor W. Zue

Room 36-541
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

[1] Paper 13 presented at the 106th Meeting of the Acoustical Society of America, San Diego, CA, November 8, 1983

Exploring Allophonic and Lexical Constraints in a Continuous Speech Recognition System. Francine R. Chen and Victor W. Zue (Room 36-545, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139)

Recent research [Zue and Shipman, JASA Vol. 71, S7; Huttenlocher and Zue, this meeting] has shown that a broad phonetic representation of speech provides strong constraints for lexical access of isolated words. In addition, Church [1983 PhD Thesis, MIT] has demonstrated the utility of detailed allophonic constraints for parsing a sentence from a phonetician's transcription. This gives reason to believe that the coupling of lexical and allophonic constraints can be a powerful tool in continuous speech recognition. The present study explores how broad phonetic constraints can be applied to a restricted continuous speech task. Using a broad phonetic representation derived from an ideal transcription, it was found that on the average, 70% of the word boundaries in the digit vocabulary can be identified. Extending this approach to speech data, we have implemented a classifier which derives a broad phonetic representation from the speech signal. Preliminary results indicate that allophonic and lexical constraints can be effective in reducing the number of string candidates, based upon the output from this classifier. [Work supported by the Office of Naval Research under contract N00014-82-K-0727 and by the System Development Foundation.]

Technical Committee: Speech Communication

Subject Classification Numbers: 43.70Sc, 43.70Ve

Telephone Number: (617) 253-7719 (F. Chen)

Send acceptance or rejection notice to F. Chen

Introduction

In a previous meeting of the Society, Zue and Shipman presented some results demonstrating the constraints imposed on the sound patterns of a language. By indexing words into a lexicon based on broad phonetic representations, the number of words sharing a common representation is very small. As a result, they proposed an approach to isolated word recognition for large vocabularies. In their proposal, the speech signal is first classified into a broad phonetic string. The broad phonetic representation is then used for lexical access, resulting in a small set of word candidates. Finally, fine phonetic distinctions are performed to determine which of these word candidates are actually spoken. We've just seen in the previous talk some refinements to their original proposal. Such an approach to isolated word recognition is in fact being pursued in a number of laboratories.

In contrast, this paper describes an attempt to exploit such constraints in continuous speech recognition. In particular, we focused our inquiries on the digit vocabulary. There are two questions which we tried to address: 1) Are there powerful enough lexical and allophonic constraints to allow words in a restricted task like continuous digit recognition to be recovered? and 2) Can such a system be realistically implemented with good performance? The first part of this talk will describe a set of experiments designed to determine whether a digit string can be recovered from an ideal transcription. Then in the second part of this talk, we will describe our first attempt to implement a speaker-independent continuous digit recognition system in which lexical candidates are reduced by using a broad phonetic classification of the speech signal.

Simulation

Constraints on sound sequences were applied first to ideal phonetic and then to ideal broad phonetic representations of digit strings to postulate word boundaries. 2000 digit strings of random order and random length containing approximately 8000 boundaries were used. We found that from an ideal, detailed phonetic transcription, every digit boundary can be positively identified. However, it is a very difficult task to automatically produce an accurate phonetic transcription across a wide population of speakers. On the other hand, we believe that producing a broad phonetic representation from the acoustic signal is not as difficult. Such a representation is also more robust against environmental and interspeaker variabilities. Thus, for the second part of this experiment, the constraints were relaxed by mapping the phones into broad phonetic categories in place of detailed phonetic transcriptions. Six broad classes were used: liquid or glide, stop, vowel, nasal, strong fricative, and weak fricative. Coarticulation effects, such as gemination of /s/ in "6-7" were ignored in producing the transcriptions.

An example of the procedure is shown in the figure. The digit string "64583", as shown on the top line, is mapped into the broad phonetic representation shown on the second line. No boundary marks are given in this broad representation, although their placement is indicated by the sharp sign shown on the line above. For this example, three of the word boundaries can be identified because no word in the lexicon contains the sequence formed by the broad labels on each side of the boundary. However, the boundary between "5" and "8" is not identified because the lexical representations of both the digits "4" and "5" contain the sequence "weak-fricative vowel". Performing this experiment on the 2000 digit strings, 70% of the word boundaries were found.

The ability to definitely identify 70% of the word boundaries is impressive, but not necessarily surprising. For example, Church described in his doctoral thesis how detailed allophonic constraints can be used to successfully parse a sentence from a phonetician's transcription. However, these results may not be directly applicable to real data. With real data, phonetic variabilities and front-end errors dictate that boundaries only be proposed. Instead of positively identifying word boundaries, a system could propose words and their corresponding boundaries by examining the "sequence" of broad class labels. That is, the word boundaries would be proposed at the beginning and end of each sequence. In addition, allophonic constraints can help to reduce the possibly large number of word candidates. A preliminary system for exploring these lexical and allophonic constraints will be described next.

Implementation

We have implemented a broad phonetic classifier and a lexical access component that will be part of a continuous digit recognition system. In this implementation, eight phonetic classes, as shown in the figure, were used. Note that they are slightly different than those used in the early part of the study. These symbols will be used throughout the remainder of this talk. The envisioned general structure of our recognition system is shown in the next figure. From the speech signal, a broad phonetic classification is first performed. The resulting broad phonetic representation is of the form of a lattice composed of broad phonetic labels. From the broad phonetic representation, the lexical access component produces a lattice of word candidates for the system. This set of candidates is a reduced set from all possible candidates and can be given to a verification component which would use more detailed acoustic analysis to identify the true sentence.

The broad classifier extracts acoustic parameters from the speech signal and characterizes the resulting parameters as acoustic features. From the set of features, the system uses a set of production rules to deduce which broad classes may be present at each segment. There are several important attributes of the classifier. One is that it begins classification by labeling regions of the speech signal which can be identified robustly. Another attribute is that in regions where the cues are not as robust, more than one label is allowed. By allowing ambiguity in the labels under uncertainty, as the classifier does, it is insured that the correct answer is not ruled out unless there is no evidence for it.

Lexical Access and Allophonic Constraints

In lexical access, lexical and allophonic constraints are used to map broad class transcriptions of the words in the lexicon to the lattice of broad labels, and indirectly, to the speech signal. The next Figure illustrates how allophonic constraints are utilized in the lexical representation of the digits. The context in which each pronunciation occurs can then be used to constrain when a word is hypothesized. In the Figure, three pronunciations of the word "8" are shown. A sample context in which the second and third pronunciations could occur is shown on the right. Each pronunciation differs in the allophone of /t/. This can be captured at the broad level as shown in the center column. A released /t/ is transcribed as a released stop. And an unreleased /t/ and flap are transcribed as silence and a short voiced obstruent, respectively.

The set of reference transcriptions that are used for lexical retrieval is produced by the system from a given set of pronunciations of the words in the lexicon. The derived transcriptions contain alternate broad phonetic pronunciations of each digit and the context under which they can occur. Each transcription is matched against the labeled segments of the broad class lattice. Thus errors made at one point in time will not affect recognition of the rest of the sentence. The next Figure illustrates application of lexical constraints on a sample digit string "5-8-6". Each box in the figure represents the position of a segment relative to the other segments in time, but does not convey any information about duration or rank. Part (a) depicts the broad segmentation produced by the classifier for the three digits. In (b), the lattice of matching words from the lexicon is shown. All the digits in starred boxes can be removed because each is a word candidate that will result in an incomplete path. The resulting lattice is shown in (c). Note that the starred "5" could also be removed by allophonic constraints. Allophonic constraints require that if a vowel follows the "5", the /v/ represented by the short voiced obstruent should not be deleted. The pronunciation of the starred "5" has /v/ deleted and is thus incorrect in this context.

The next figure illustrates actual constraint application in lexical access for a longer digit string. A spectrogram of the digit string "7620085", the corresponding broad representation, and the corresponding word lattice are shown. It can be observed that the word lattice is much reduced from the general case where all the words in the lexicon can begin at each segment. But it should also be noted that "3", "4", and "5" are not distinguished in the broad representation. However, a simple check, such as vowel height, may be able to differentiate among some of them.

Results

As mentioned earlier, we have implemented on a Lisp machine workstation such a speaker-independent continuous digit recognition system up to the level of lexical candidate reduction. The system was developed using the speech data from one male speaker. We have just evaluated the performance of the system for the first time last week using some two hundred digits spoken by three new speakers, one male and two female. The preliminary results indicate that 1% of the time the correct digit is not one of the lexical candidates. The corresponding depth of the digit lattice is four. We would like to stress that the system is under active development and that the performance results are very preliminary. Nevertheless, we are encouraged by the results, and feel that this may be a viable approach to continuous digit recognition.

Summary

In summary, this paper proposed a new approach to continuous digit recognition. In this approach, the speech signal is initially classified into broad phonetic categories. It was shown that allophonic and lexical constraints can be powerful even at a broad phonetic level. Thus, lexical access based on a broad phonetic description will hopefully result in a small number of candidate digit strings. Whenever more than one digit candidate spans a given time interval, fine phonetic distinctions will be performed to select the best candidate. Because of the fact that fine phonetic classification is not performed at the onset, the system has the potential of being robust against inter-speaker variabilities.

Exploring Allophonic and Lexical Constraints in a Continuous Speech Recognition System

Francine R. Chen
Victor W. Zue

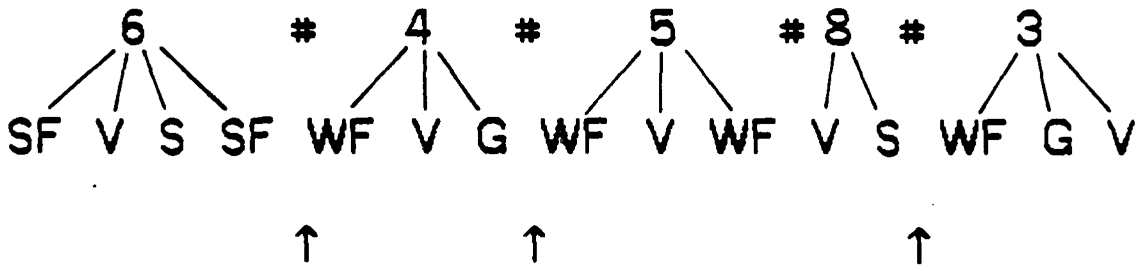
Department of Electrical Engineering & Computer Science
Research Laboratory of Electronics
Massachusetts Institute of Technology

QUESTIONS

- Are there powerful enough lexical and allophonic constraints to allow words in a restricted task like continuous digit recognition to be recovered?

- Can such a system be realistically implemented with good performance?

SEQUENCE CONSTRAINT EXAMPLE



6 # 4 # 5 # 8 # 3

SF V S SF WF V G WF V WF V S WF G V

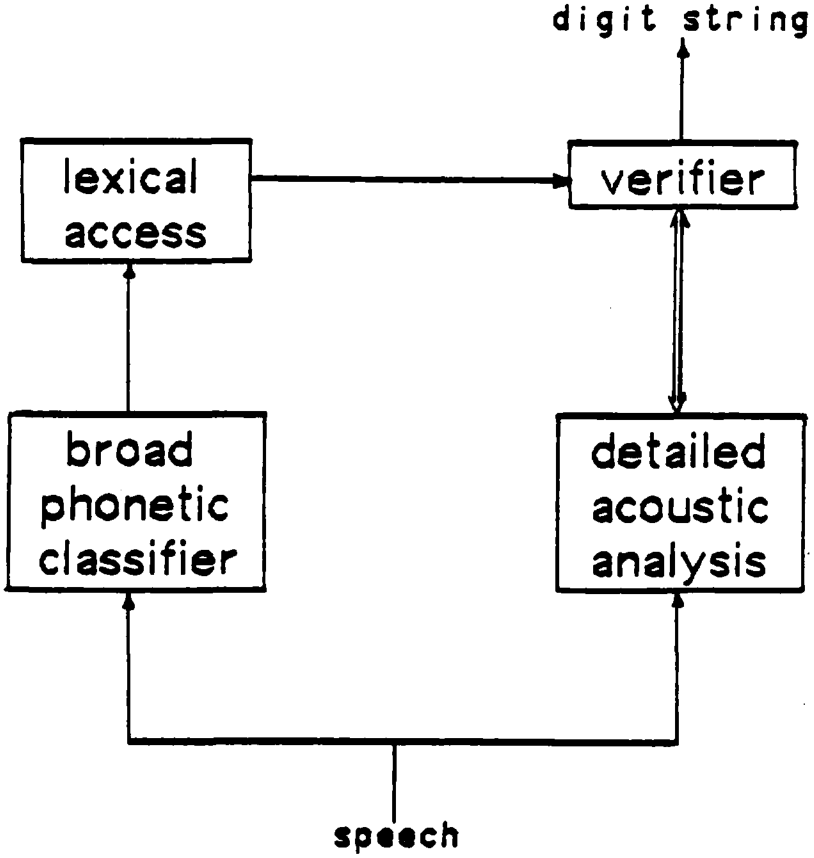
↑ ↑ ↑

G	liquid or glide	N	nasal
S	stop	SF	strong fricative
V	vowel	WF	weak fricative

BROAD CLASSES USED WITH NATURAL SPEECH

G	sonorants
N	intervocalic nasal
S	stop
SF	strong fricative
SVO	short voiced obstruent
V	vowel
WF	weak fricative
-	silence

BLOCK DIAGRAM OF A RECOGNITION SYSTEM



BROAD PHONETIC CLASSIFIER

- Produces broad class representation of input speech signal

- Principles

  *Label regions which can be identified robustly

  *Allow ambiguity when uncertain

ALLOPHONIC CONSTRAINTS

Lexical items can be represented in terms of multiple pronunciations

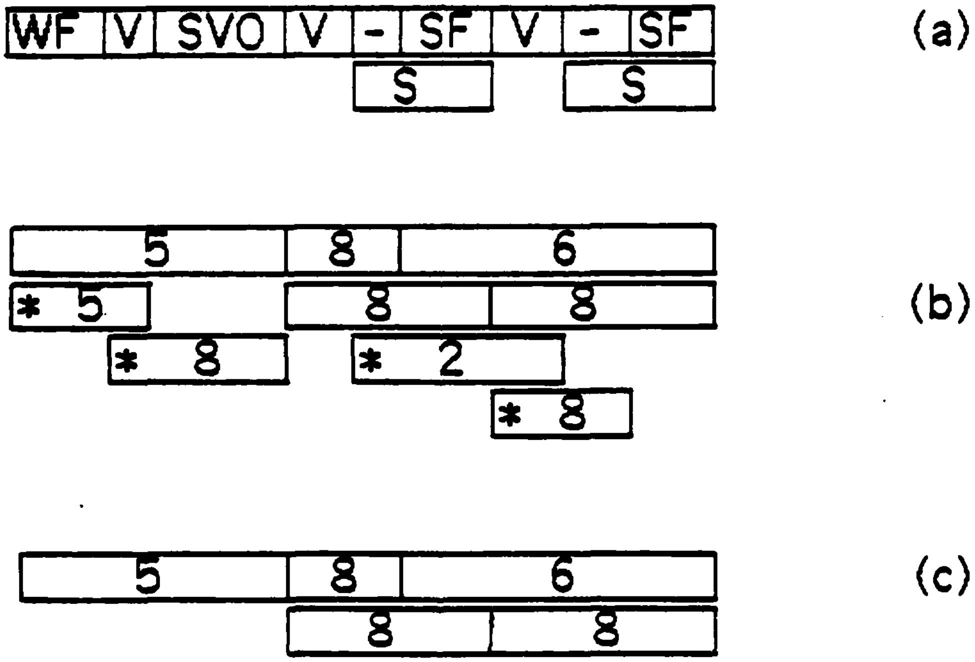
Example:

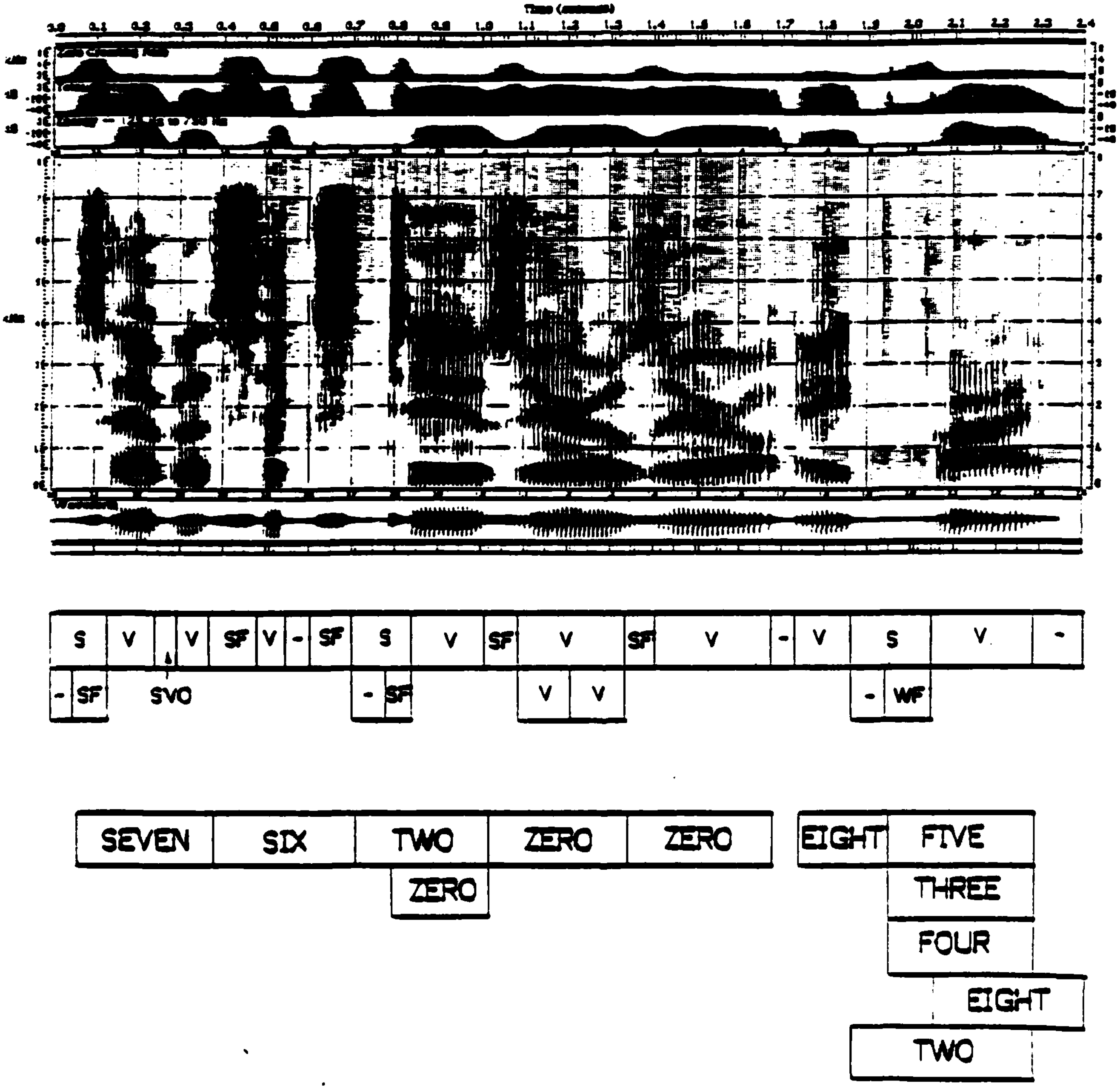
$e^{y}t$	vowel stop	
$e^{y}t^{\square}$	vowel silence	/_ nasal
e^{y}ɾ	vowel svo	/_ vowel

LEXICAL CONSTRAINTS

Eliminate word candidates that will result in an incomplete path

Example:




Time (seconds)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4
Zero Crossing Rate
Energy
Energy -- 125 Hz to 750 Hz
kHz
Waveform
S V V SF V - SF S V SF V SF V - V S V -
- SF SVO - SF V V - WF
SEVEN SIX TWO ZERO ZERO EIGHT FIVE
ZERO
THREE
FOUR
EIGHT
TWO

SUMMARY

- Allophonic and lexical constraints can be powerful even at a broad phonetic level for digit recognition

- Lexical access based on broad phonetic information can help to reduce the number of digit candidates

- Such a system can potentially be speaker-independent

The Use of Structural Constraints to Determine Word Boundaries[1]

by
Lori F. Lamel

Room 36-545
Department of Electrical Engineering and Computer Science
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

[1] Paper I3 presented at the 106th Meeting of the Acoustical Society of America, San Diego, CA, November 8, 1983

The Use of Phonotactic Constraints to Determine Word Boundaries. Lori F. Lamel (Room 36-545, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139)

Phonotactic constraints limit the permissible word internal consonant sequences in English. In some cases, knowledge of the phoneme sequence uniquely specifies the location of the word boundary, while in other cases, phonological rules based on allowable consonant sequences are not sufficient. For example, the word boundary can be uniquely placed in the sequence /... m g l ... /, as in the word pair "some glass", whereas the the word boundary location is ambiguous in the phoneme sequence /... s t r ... / without further acoustic information. The /... s t r ... / may have a word boundary in one of three places as in "last rain", "race trials", and "may stretch". Studies were conducted to determine the utility of phonotactic constraints to predict word boundaries. The databases included the Merriam-Webster Pocket Dictionary, a phonemically balanced set of sentences, and samples of unrestricted text. Results indicate that: 1) word internal consonant sequences represent a very small subset of all permissible consonant sequences across word boundaries; and 2) acoustic-phonetic knowledge is needed when the word boundary is ambiguous. Results on the differences between word and syllable boundaries will also be presented. [Work supported by the Office of Naval Research under contract N00014-82-k-0727 and the System Development Foundation.]

Technical Committee: Speech Communication
Subject Classification numbers: 43.70.Sc,43.70.Ve
Telephone number: (617)-253-7719

This study investigates the occurrence of consonant sequences at word boundaries and within words in English. The main question raised is "Given a consonant sequence, can a word boundary location be determined?". Earlier work [Zue and Shipman, J. Acoust. Soc. Am. Suppl. 1 71,S7 (1982)] has shown that there are lexical constraints limiting the number of possible consonant sequences within words. The first problem that we address is what, if any, structural constraints on consonant sequences in English limit the potential word boundary sequences.

The second problem we address is whether there are acoustic cues to word boundaries in continuous speech, and how these could be used for speech recognition. Listeners are able to hear individual words even though words in continuous speech are not separated by pauses. Nakatani and Dukes (1979) have shown that word boundaries, as well as syllable boundaries, are often marked by acoustic cues that listeners use for speech perception. We believe that a good understanding of the acoustic manifestations of phonemes at word boundaries will enable us to resolve phonetic ambiguity and to propose potential word boundaries from acoustic evidence.

To address the first problem, we determined the occurrences of consonant sequences in a corpus of text files. The text files ranged in size from 200 words to 38,000 words and were obtained from a variety of sources. The phonemic transcription of each word was obtained by dictionary lookup.

The number of distinct consonant sequences were determined for within words and across word boundaries. The number of occurrences of each distinct sequence was also recorded. Word boundary sequences were formed by concatenating the word-final consonant sequence of the current word with the word-initial consonant sequence of the next word. In this example, the word pair "western front" has a word boundary sequence /nfr/, whereas the word "western" has the medial sequence /st/. There are approximately 70 distinct word-initial consonant sequences, and 130 word-final sequences. Based on these estimates there are potentially over 9000 word boundary consonant sequences. The top curve in the figure shows the upper bound on the number of possible word boundary sequences as a function of the number of words in the corpus. The upper bound arises by assuming that any word can follow any other word, and is thus the product of the number of word-final and word-initial consonant sequences for the given corpus.

The lower curve shows the number of word boundary sequences occurring in the corpus. In general, this curve is about 25% of the upper bound. The behavior of the two curves shown here are similar. We can expect that perhaps the lower curve will approach the upper bounding curve as larger samples of text are processed. However, if there are structural constraints limiting the combinations of words in English, then perhaps the consonant sequences that occur across word boundaries will remain a subset of the total number of possibilities.

The overlap between the word-medial sequences and the word-boundary sequences gives a comparison of what goes on within words and across word boundaries. If the lexical constraints on word internal sequences and the structural constraints across word boundaries are the same, then the curves for the two conditions should be similar. The overlay shows the number of word-medial consonant sequences. On average only 20% of the distinct word boundary sequences occur in word medial position. This means if I were to randomly choose a sequence from the distinct consonant sequences, then, on the average, 4 out of 5 times the sequence would specify a word boundary which would not be a syllable boundary. If I instead put all the word boundary sequences, occurring as many times as they do in the corpus, into a bucket, then 1 out of 3 randomly chosen sequences will only occur at a word boundary. [When the frequencies of occurrence are accounted for, 1 out of 3 consonant sequences specifies a word boundary.]

There is also a difference in the average lengths of consonant sequences in word-medial and word-boundary position. Word medial sequences are shorter than word boundary sequences. The average length for sequences within words is 2.2, while across word boundaries it is 2.9. As will be discussed next, longer sequences have more constraints imposed upon them.

How many of the distinct consonant sequences have a unique boundary location? A unique boundary location means that the consonant sequence can only be divided in one way such that the sequence forms an allowable offset cluster followed by an allowable onset cluster. For example, the sequence /mgl/ as in "some glass" can only have a word boundary between the /m/ and the /g/. The phoneme sequence /sts/ can form an allowable offset as in "casts" or occur across a word boundary as is "last side". About 65% of the word boundary sequences have a unique boundary location. Another 30% have 2 possible locations. The word boundary sequences not occurring in medial position have a lower rate of ambiguity, with approximately 80% having the boundary location uniquely specified. Part of these differences can be accounted for by considering the length of sequences in the two positions. As mentioned before, on average, word boundary sequences are almost 1 consonant longer than medial sequences. The average number of boundary locations decreases with increasing length of the consonant sequence. 60% of the consonant sequences of length 2 have ambiguous boundary locations while less than 20% of longer sequences have ambiguous boundary locations.

Given that an ideal phonemic transcription cannot uniquely specify a boundary location in one third of the consonant sequences occurring at word boundaries, can these boundaries be disambiguated? Our belief is that in some cases acoustic evidence may be useful. To this end we conducted a preliminary experiment to investigate acoustic cues to word boundaries in labial-stop sonorant clusters. Minimal pair phrases such as "grape lane" and "grey plane", as shown in the figure, were embedded in a carrier phrase and recorded by 3 male speakers. The recordings were digitized and analyzed using SPIRE on the Lisp Machine. Parameters were extracted from the transcribed utterances and statistical analysis performed.

The spectrograms of "grape lane" and "grey plane" have several differences. These include the duration of the first vowel in each pair, the amount of aspiration in the /p/, VOT, and some characteristics of the /l/, such as a steady state region and formant frequencies at voice onset.

Some simple measures that can be used to differentiate this pair are the duration of the release portion of the stop, the duration of the sonorant, and the formant frequencies at voice onset. Shown in black is a histogram of the stop release duration for /p/ in word-initial /pl/ clusters. The devoiced portion of the /l/ is included as part of the release as was shown in the previous figure. A histogram of the duration of the release of /p/ in /p#l/ is shown in red. All unreleased /p/'s were given a duration of 0 ms. As can be seen, about half of the time the /p/ was unreleased, and in the remainder there is a short, generally weak, release.

This figure shows a scatter plot of the second and third formant at voice onset for the /r/ in initial /br/ clusters and for /b#r/. Although there still a fair amount of overlap between the two conditions, the combination of F_2 and F_3 give better discrimination than either alone.

Throughout this talk we have intentionally avoided the topic of syllable boundaries, and the comparison of syllable and word boundaries. There are several reasons for this. First, we do not know what the correct syllabification of words should be. We have looked at two different syllabifications of a large lexicon, but neither shows the consistency we would like to have. In the literature there are various approaches to syllabification and we are looking into this issue further. It seems that we have a bit of the "chicken and the egg" problem here. We would like to have some theory for syllabification which could be used to predict acoustic manifestations of consonant sequences at syllable boundaries. However, we believe that acoustic evidence is necessary to get reliable syllabification.

In summary, we have found that there are structural constraints on the allowable sequences of consonants at word boundaries in English. 4 out of 5 word boundary sequences do not occur within a word. Of these sequences 80% have a word boundary which is uniquely specified by the sequence of phonemes. In the cases where the phoneme sequence cannot uniquely specify the word boundary acoustic information may be of help.

The Use of Structural Constraints
to Determine Word Boundaries

Lori F. Lamel

Room 36-545
Department of Electrical Engineering & Computer Science
Research Laboratory of Electronics
Massachusetts Institute of Technology

CONSONANT SEQUENCES AT WORD BOUNDARIES

- Are structural constraints useful to delimit word boundaries?

- Are there acoustic cues to word boundaries?

STUDY OF STRUCTURAL CONSTRAINTS BASED ON IDEAL TRANSCRIPTION

CORPUS

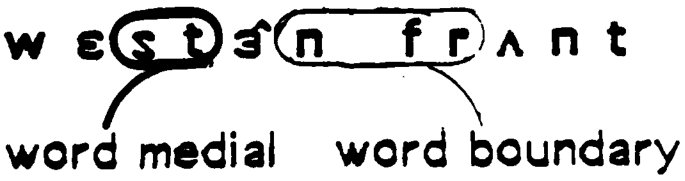
- Text files ranging in size from 200 words to 38,000 words

PROCEDURE

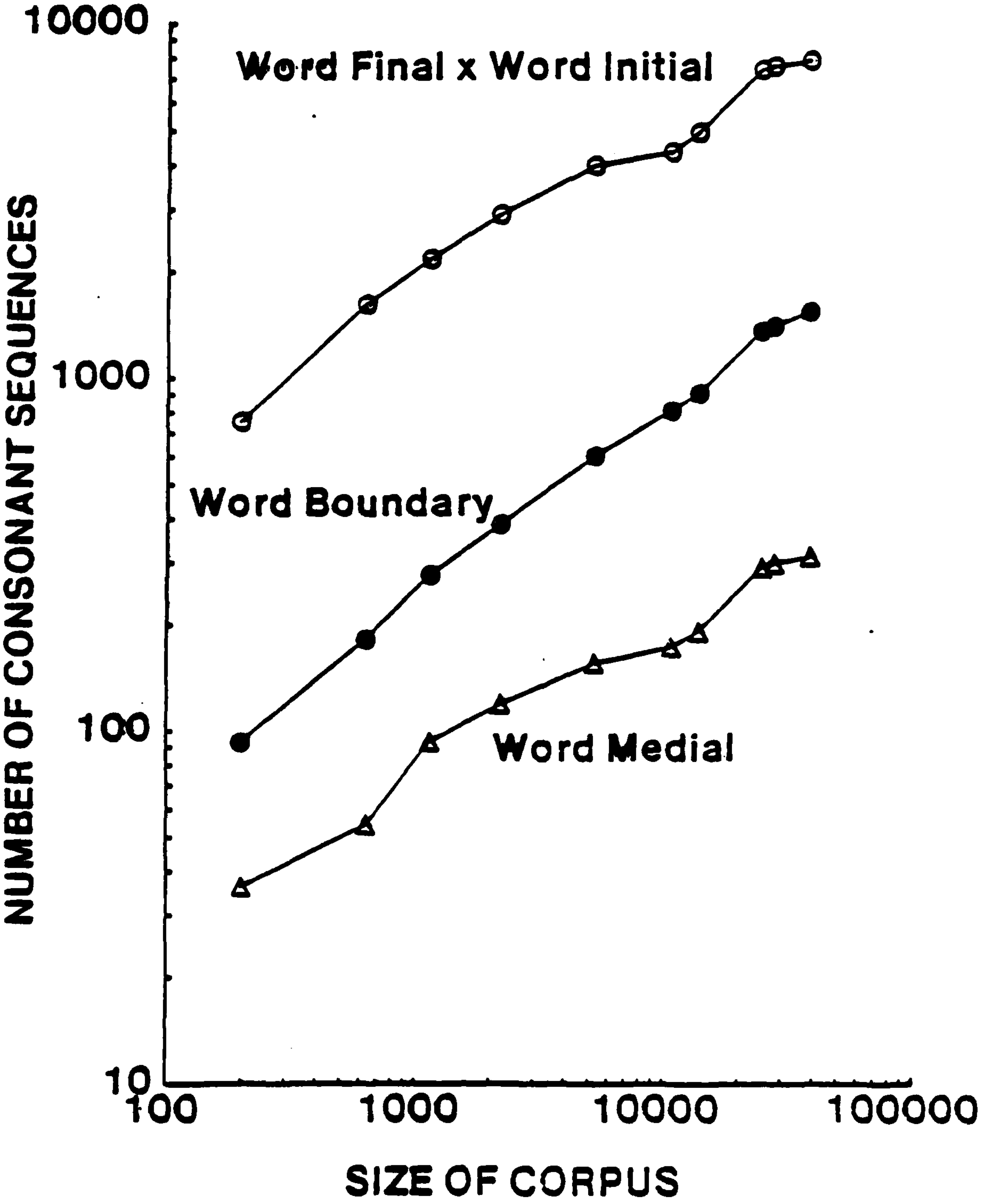
- Identify distinct consonant sequences within words and across word boundaries

- Record the number of occurrences for each distinct consonant sequence

WESTERN FRONT


w ɛ s t ɝ n f r ʌ n t

word medial word boundary


Word Final x Word Initial
Word Boundary
Word Medial
NUMBER OF CONSONANT SEQUENCES
10000
1000
100
10
100
1000
10000
100000
SIZE OF CORPUS

WHAT IS A UNIQUE BOUNDARY LOCATION?

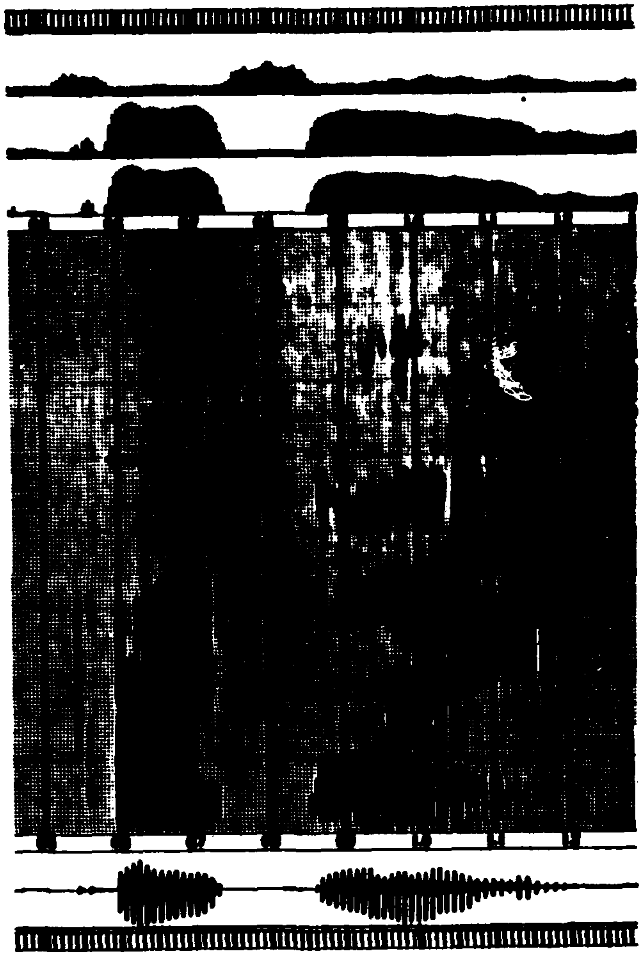
CONSONANT SEQUENCE ==> Permissible coda followed by permissible onset

EXAMPLES:

/mgl/	==>	m # gl	"some glass"
/sts/	==>	sts #	"casts"
		st # s	"last side"

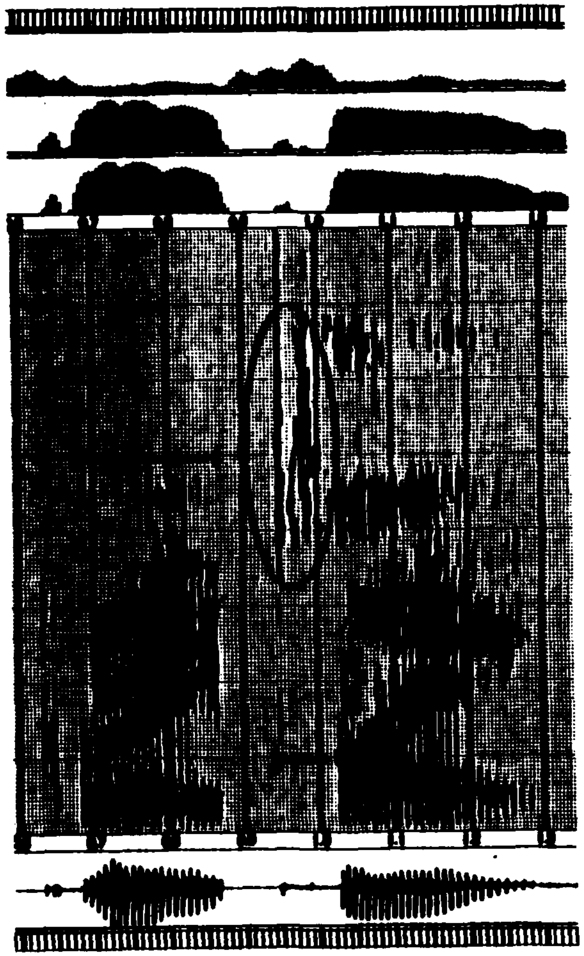
HOW MANY CONSONANT SEQUENCES HAVE A UNIQUE BOUNDARY LOCATION?

- 65 % of word boundary sequences have unique locations
- 80 % of word boundary sequences not occurring in word medial position have unique boundary locations
- Most ambiguous sequences have only two possible boundary locations



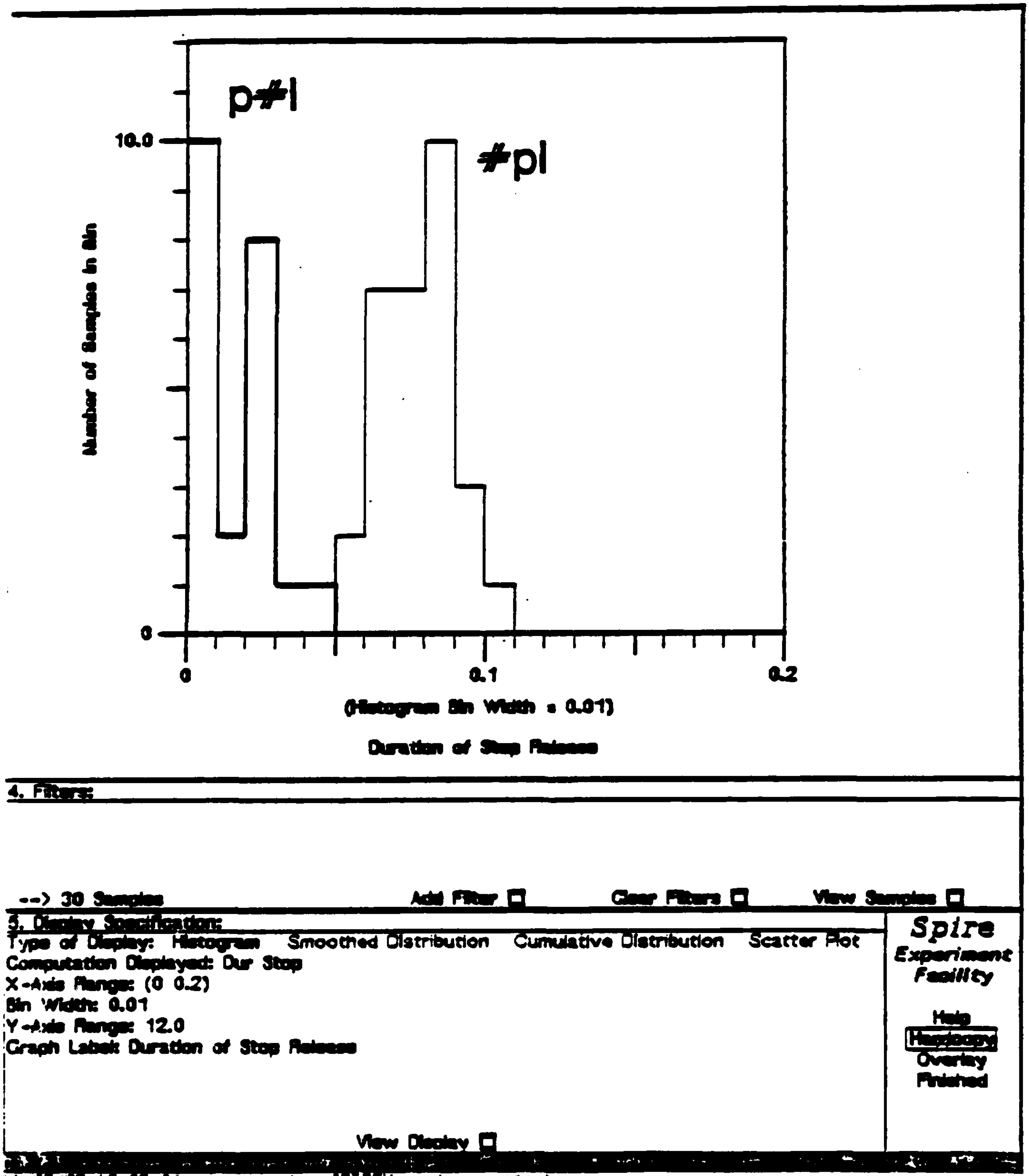
g r e^{y} p$^{\urcorner}$ l e^{y} n

grape lane



g r e^{y} p l e^{y} n

grey plane



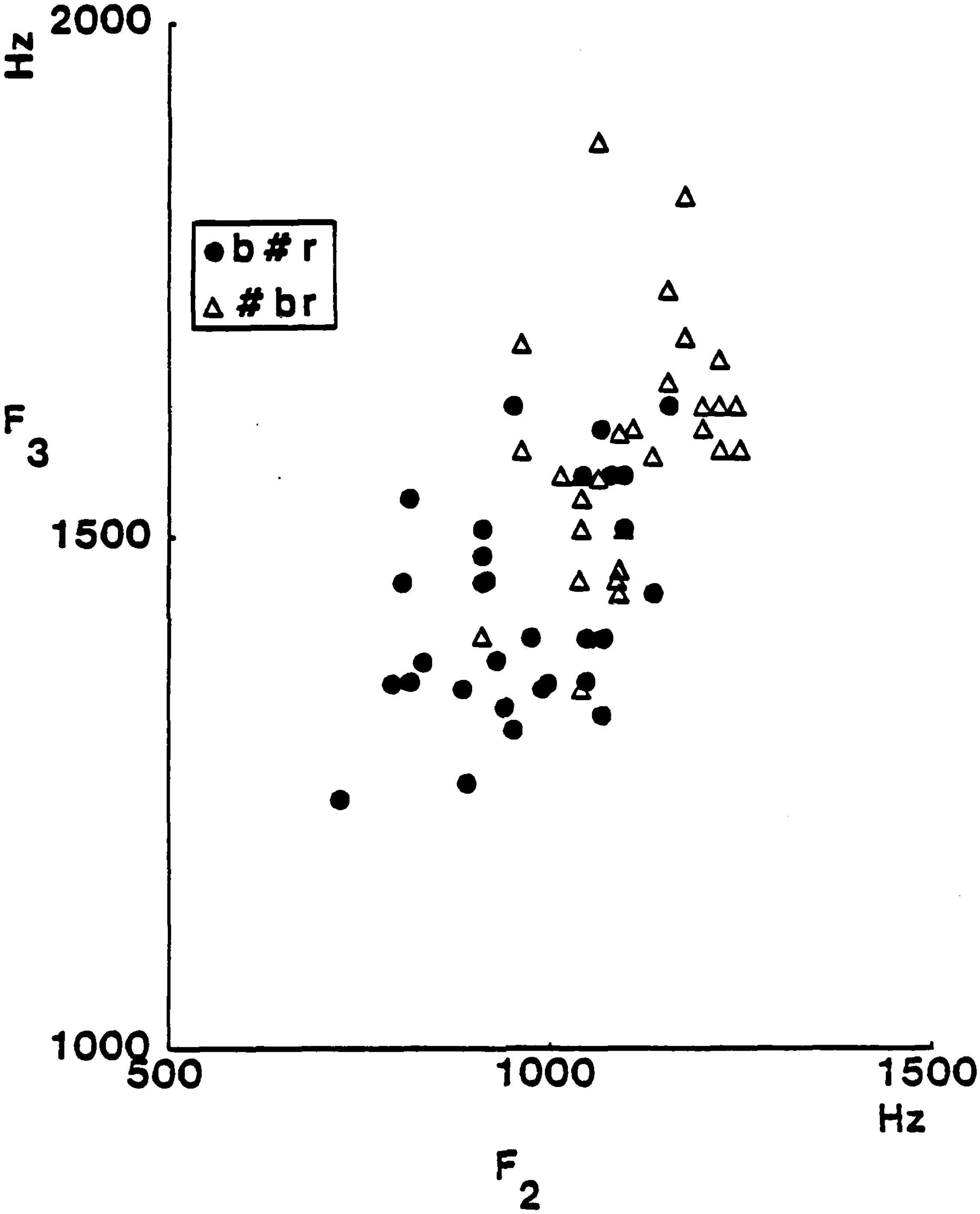
4. Filters:

--> 30 Samples

5. Display Specification:

Type of Display: Histogram Smoothed Distribution Cumulative Distribution Scatter Plot
Computation Displayed: Dur Stop
X-Axis Range: (0 0.2)
Bin Width: 0.01
Y-Axis Range: 12.0
Graph Label: Duration of Stop Release

?/02/83 18:23:34 Lori Lamel SPIREX: (no_window)



Hz
2000
1500
1000
F3
● b # r
△ # b r
500
1000
1500
Hz
F2

SUMMARY

- 4 out of 5 consonant sequences occur only at word boundaries

- 80% of word boundary sequences have only one possible boundary location

- Acoustic information may help to locate word boundaries

presented at the International Conference on
Advanced Automation - 1983, Taiwan, Republic of China
December, 1983

COMPUTER RECOGNITION OF ISOLATED WORDS FROM LARGE VOCABULARIES: LEXICAL ACCESS USING PARTIAL PHONETIC INFORMATION

Victor W. Zue and Daniel P. Huttenlocher[*]
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, Massachusetts 02139 U.S.A.

Current approaches to isolated word recognition rely on classical pattern recognition techniques which utilize little or no speech specific knowledge. While the performance of these systems is quite good for a set of restricted vocabularies and tasks, they are not readily extendible to more complex tasks. This paper proposes a new approach to isolated word recognition intended for multiple speakers and large vocabularies. The system performs a broad phonetic categorization of the acoustic signal. This partial phonetic information is then used for lexical access, generating a small set of possible word candidates. Fine phonetic distinctions can then be performed in order to determine which of these word candidates was actually spoken. Preliminary results on lexical access and word candidate reduction is presented.

1. INTRODUCTION

During the past decade, significant advances have been made in the field of isolated word recognition (IWR). Today, speech recognition systems can typically recognize a small set of words, say 50, for a given speaker with an error rate of less than 5%. Current approaches generally use a pattern matching technique where the input signal is matched against a set of stored templates. The success of these systems can at least in part be attributed to the introduction of novel parametric representations[1], distance metrics[2], and the very powerful time alignment procedure called dynamic programming[3].

While we have clearly made significant advances in dealing with the IWR problem, there is serious doubt regarding the extendibility of the pattern matching approach to large vocabulary, speaker-independent isolated word recognition. One of the limitations of the template matching approach is that both storage and computation grow linearly with the size of the vocabulary. That is, a separate template is stored for each word, and recognition involves comparing each template with the unknown utterance. For very large vocabularies (several thousand words or more), these computation and storage requirements become prohibitively expensive. Even if computational cost were not an issue, the performance of such IWR systems deteriorates for large vocabularies[4]. Furthermore, as the size of the vocabulary grows, it becomes imperative that recognition systems be speaker-independent, since training the system for each speaker will be painfully impractical.

In this paper, we propose a new approach to large-vocabulary, isolated word recognition. This approach relies on partial phonetic information for lexical access, thereby greatly reduces the number of word candidates. This partial phonetic representation derives its power from the high degree of redundancy in a given language. By exploiting this redundancy, it is possible to eliminate all but a handful of words out of a 20,000-word lexicon, using only half a dozen broad phonetic classes. In the next section of the paper, we summarize a set of studies demonstrating the extent to which a partial phonetic representation can be used to eliminate all but a small set of word candidates. The subsequent section describes the system being implemented, and presents some preliminary results.

2. DESIGN PHILOSOPHY

2.1 Spectrogram Reading

Reliance on general pattern matching techniques has been partly motivated by the unsatisfactory performance of early phonetically-based speech recognition systems. In fact, the difficulty of the acoustic-phonetic recognition task has led to speculation that phonetic information must be derived primarily from syntactic, semantic and discourse constraints rather than from the acoustic signal. However, the poor performance of early phonetically-based recognition systems can be attributed mainly to our limited knowledge of the acoustic characteristics of speech sounds, particularly the effects of local context. This picture is slowly changing. We now have a far better understanding of contextual influences on phonetic segments, as evidenced by a series of spectrogram reading experiments[5]. It was found that a trained subject can phonetically transcribe unknown utterances from speech spectrograms with an accuracy of approximately 85%. This level of performance is far better than the phonetic recognizers reported in the literature, both in terms of accuracy and rank order statistics. It was also demonstrated that the process of spectrogram reading makes use of explicit acoustic phonetic rules, and that this skill can be learned by others. These results suggest that the acoustic signal is rich in phonetic information, and that it may be possible to obtain substantially better performance in automatic phonetic recognition.

Even with substantially improved acoustic-phonetic knowledge however, an approach based entirely on detailed phonetic analysis still has serious drawbacks. Using solely acoustic information, it is often difficult to make fine phonetic distinctions. (For example, it is difficult to distinguish the word pair "Sue/shoe" reliably across a

[*] Research supported by the Office of Naval Research under Contract N00014-82-K-0727 and by the System Development Foundation

wide range of speakers). Furthermore, the application of context-dependent rules requires the specification of the correct context. (For example, the identification of a retroflexed /t/ in the word "tree" depends upon correctly identifying the retroflex consonant /r/). Thus, recognition based solely on *detailed* phonetic transcription of an unknown utterance may not be desirable, or even possible.

2.2 Constraints on Sound Patterns

The sound patterns of a given language are not only limited by the inventory of basic sound units, but also by the allowable combinations of these sound units. Knowledge about such phonotactic constraints is presumably used in speech communication, since it provides native speakers with the ability to *fill in* phonetic details that are otherwise not available or are distorted. Thus, as an extreme example, a word such as "splint" can be recognized without having to specify the *detailed* acoustic characteristics of any phoneme other than the /l/, because it is the *only* word in the Merriam Webster's Pocket Dictionary (containing about 20,000 words) that satisfies the following description:

[CONSONANT][CONSONANT][l][VOWEL][NASAL][STOP]

While the existence of phonotactic constraints is well known, a recent set of studies[6,7] provides a glimpse of the magnitude of their predictive power. These studies examine the phonotactic constraints of American English from the phonemic distributions in the 20,000-word Merriam Webster's Pocket Dictionary. In one study the phonemes of each word were mapped into one of six broad phonetic categories: vowels, stops, nasals, liquids and glides, strong fricatives, and weak fricatives. Thus, for example, the word "speak", with a phonemic string given by /spik/, is represented as the pattern:

[STRONG FRICATIVE][STOP][VOWEL][STOP]

It was found that, even at this broad phonetic level, approximately 1/3 of the words in the 20,000-word lexicon can be uniquely specified. One can view the broad phonetic classifications as partitioning the lexicon into equivalence classes of words sharing the same phonetic class pattern (e.g., the words "speak" and "steep" are in the same equivalence class). The average size of these equivalence classes for the 20,000-word lexicon was found to be approximately 2, and the maximum size was approximately 200. In other words, in the worst case, a broad phonetic representation of the words in a large lexicon reduces the number of possible word candidates to about 1% of the lexicon. Furthermore, over half of the lexical items belong to equivalence classes of size 5 or less.

2.3 Dealing with Phonetic Variability

The results of the Shipman and Zue study demonstrate that broad phonetic classifications of words can, in principle, reduce the number of word candidates significantly. However, the acoustic realization of a phone can be highly variable, and this variability introduces a good deal of ambiguity in the initial classification of the speech signal. At one extreme, the acoustic characteristics of phonemes can can undergo simple modifications as a consequence of contextual and interspeaker differences, such as the differences in the acoustic signal for the various *allophones* of /t/ in words "tea", "tree", and "beauty". At the other extreme, contextual effects can also produce severe modifications in which phonemes or syllables are deleted altogether. Thus, for example, the word "international" can have many pronunciations including the deletion of the phoneme /t/ and the deletion of the penultimate schwa.

It is important to note that these phenomena can occur as a consequence of the high level of variability in natural speech, as well as resulting from an error by the front-end classifier of a speech recognition system. Given such uncertainties, one may ask whether the original results of Shipman and Zue still hold for lexical access. The answer is partially provided in a recent study conducted by Huttenlocher and Zue[7], in which they observed the effects on lexical constraints after introducing various amounts of phonetic uncertainty into the lexicon. They found that, even allowing as much as 20% phonetic uncertainty, lexical constraints imposed by sequences of broad phonetic classes are still extremely powerful. Over 30% of the lexical items can still be uniquely specified, and over 50% of the time the size of the equivalence class is 5 or less. On the other hand, the *maximum* sizes of the equivalence classes grow steadily as the amount of labeling uncertainty increases.

3. SYSTEM DESCRIPTION

3.1 Overview

Based on the results of the studies cited above, we propose a new phonetically-based approach to isolated-word recognition. This approach is distinctly different from previous attempts in that detailed phonetic analysis of the acoustic signal is not performed. Rather, the speech signal is classified into several broad manner categories which are then used directly for lexical access. The broad phonetic (manner) classifier serves several purposes. First, errors in phonetic labeling, which are most often caused by detailed phonetic analyses, should be reduced. Second, by avoiding fine phonetic distinctions, the system should also be less sensitive to interspeaker variations. Finally, experimental results indicate that, even at the broad phonetic level, sequential constraints and the lexical distribution can limit the search space substantially. This last feature is particularly important when the size of the vocabulary is large (on the order of several thousand words or more).

Once the acoustic signal is reduced to a sequence (or lattice) of broad phonetic segments, the resulting representation is used for lexical access. The intent is to reduce the possible word candidates to a very small set by utilizing knowledge about the structural constraints, both segmental and suprasegmental, of the words. Lexical access is performed by indexing into an "inverted lexicon", where each word is stored in terms of its broad phonetic classification. Since the broad phonetic classifications specify such small subsets of the lexicon, the resulting set of word candidates should be very small. From this word set, the correct word can then be selected through the judicious application of detailed phonetic knowledge.

We should point out that the system uses a very conservative control strategy. Assertions about the acoustic or phonetic identity of a speech segment are only made when the evidence for that classification is very strong. In this manner, a reliable gross acoustic description is obtained. Detailed classification is left until after lexical access, when specific phonetic hypotheses can be formed and evaluated. This approach can also be viewed as a late binding approach, in which binding a segment to a label (or set of labels) is delayed as long as possible. A block diagram of the processing performed by the phonetic class recognizer is presented in *Figure 1. The remainder of this section will follow the* outline of the block diagram.

WAVEFORM
↓
PARAMETERIZATION
↓
CHARACTERIZATION
↓
ACOUSTIC SEGMENTATION
↓
PHONETIC-CLASS SEGMENTATION
↓
LEXICAL ACCESS

Figure 1: Block Diagram of Broad Phonetic Classification

3.2 Signal Parameterization

The parameterization stage consists of extracting a set of parameters from the acoustic waveform. The speech signal is sampled at 16 kHz and passed through a filter bank. The energy in each band is then calculated every 5 msec, using a 25 msec window. In general, the short-time variations in these energy contours are overly detailed, since many of the small changes in energy are irrelevant to the broad phonetic identification task. Ideally, we would like to preserve only that acoustic information which is relevant to broad phonetic events. However, simply smoothing the contour over a long time window is not appropriate, because some short-term events are important. Therefore, the parameterization stage of processing is designed to remove the irrelevant information in the energy parameters while preserving that information needed for phonetic identification. This is done by producing a stepwise approximation to the energy contour, where the steps correspond to longer term acoustic events. The approximation preserves the area of the original energy contour by making each step the same area as the portion of the curve within that step. Recent work on scale-space filtering is also concerned *with producing qualitative descriptions of signals. However, that* work is concerned with forming such descriptions independent of the process underlying the signal[6]. Figure 2 contains some energy contours together with their stepwise approximations.

3.3 Characterizing the Parameters

The next stage of processing is concerned with producing symbolic characterizations of the stepwise approximations to the energy contours. The stepwise approximations preserve two kinds of information: magnitude and relative magnitude. Similarly, the symbolic characterization of each step is in terms of both magnitude and relative magnitude. The magnitude of each step is either LOW, MEDIUM or HIGH, and the relative magnitude is either PEAK, VALLEY, ONSET, or OFFSET. These magnitude and relative magnitude characterizations are formed by taking advantage of the natural ordering and distribution of the segments, rather than using arbitrary thresholds. For example, a step is said to be a PEAK if its magnitude is greater than that of both its neighbors. Figure 3 presents a stepwise approximation and its categorizations.

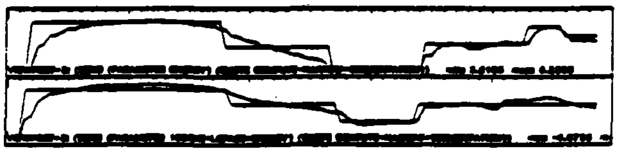


Figure 2: Energy Contours and Their Stepwise Approximations for the Word "length"

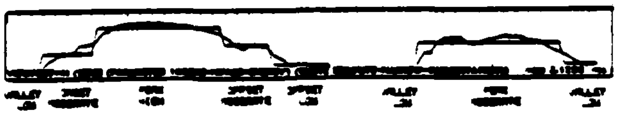


Figure 3: Categorization of the an Energy Contour for the Word "length"

3.4 Acoustic Segmentation

Given the characterizations of the energy in various frequency bands, we need some way of combining this information in order to produce an acoustic segmentation of the speech signal. We have chosen to combine information using predicates on the bandpass energy characterizations. For example (LOCAL-PEAK TOTAL-ENERGY) indicates where the TOTAL-ENERGY is a PEAK with respect to the neighboring steps. Similarly (HIGH-AMPLITUDE TOTAL-ENERGY) specifies where the TOTAL-ENERGY is HIGH in magnitude. Acoustic classes are then defined in terms of simple combinations of these predicates. For example, let us consider a potential rule for strongly voiced segments (such as strong vowels):

```
(DEFINE-ACOUSTIC-CLASS  STRONGLY-VOICED
  (DURATION-BETWEEN  30.  NIL
    (AND (HIGH-AMPLITUDE  LOW-fREQUENCY-ENERGY)
         (LOCAL-PEAK  TOTAL-ENERGY)
         (LOCAL-PEAK  LOW-FREQUENCY-ENERGY))))
```

This rule states that a strongly voiced segment must have a lot of low-frequency energy, and that both total-energy and low-frequency energy must be local peaks. The rule also specifies that the acoustic segment be at least 30 milliseconds long (with no limit on the maximum duration).

The control structure of the acoustic classifier is basically that of a simple production system, where zero or more rules may fire at any point in time. That is, each rule is triggered when the input matches its conditions. This kind of "free response" control structure has the advantage of not forcing a segment label at every point in time, nor requiring the start and end of successive labels to match exactly. As we noted above, this is a late binding strategy which puts off labelling a segment until there is good evidence for that label. Thus, the output of this level is a set of acoustic labels which can be overlapping or have gaps between them. It may seem that allowing overlap and unaccounted for gaps will cause problems in recognition. However, lexical access is performed using sequences of phonetic classes. In the end, what we care about is having a phonetic-class sequence which accurately preserves the identity and order of the phones present in the acoustic signal. We do not, in general, care exactly where in time the classes start and stop. In addition, forcing labels at each point in time can erroneously introduce false segments on the basis of poor phonetic information. These "false segment" errors will produce a phonetic-class sequence which is incorrect.

3.5 Phonetic Classification and Lexical Access

The final stage of the broad phonetic classification uses local acoustic context to produce a sequence of phonetic class labels. This is done using a rewrite grammar, which maps acoustic segments and contexts onto corresponding phonetic segments. For example, the rule for aspirated stop consonants maps silence followed by weak turbulence to an aspirated stop. The output of this stage is a sequence (or lattice in the case of multiple possible classifications) of broad phonetic classes. The broad phonetic class sequence is then used to index into the 20,000-word lexicon. The lexicon is stored in inverted form, with each entry hashed according to its broad phonetic classification. Therefore, a single hash lookup produces the set of words matching a given sequence of broad phonetic classes. In the case of a segment lattice, a small number of hash lookups must be performed. However, the occurrence of lattices is limited to the case where two broad phonetic classes cannot be reliably differentiated, for instance strong fricatives versus affricates in utterance initial position. Figure 4 contains the broad acoustic and phonetic classifications for the word "length" uttered by a male speaker, together with the lexical entries which match the phonetic class sequence.

4. RESULTS AND SUMMARY

A version of the system proposed in this paper has been implemented on a Lisp Machine based workstation in our laboratory. Preliminary results, based on 100 word tokens spoken by two male talkers, are encouraging. Over 85% of the time, the correct word was in the set of word candidates. For example, the word "century" spoken by a male speaker resulted in 4 word candidates as follows:

central	sɛntrəl
century	sɛnčərí
sensory	sɛnsərí
sentry	sɛntrí

whereas the same word spoken by another speaker resulted in 11 wnrd candidates. In both of these examples, the correct word is one of the word candidates. We are continually refining the rules for acoustic segmentation and phonetic classification. Our goal for the initial lexical access is to reduce the number of word candidates to less than 10 on the average, with an error rate of less than 5%.

```
Acoustic Classification:
```

STRONGLY-VOICED VOICEBAR SILENCE WEAK-TURBULENCE

```
Phonetic Classification
   (optional segments in parens):
```

(VOICED-STOP) (VOICED) STRONG-VOWEL NASAL WEAK-FRIC

```
Lexical Entries from 20,000-Word Lexicon:
```

8 entries:

amaranth	æmərænθ
length	lɛŋθ
lymph	lɪmf
month	mʌnθ
ninth	na'nθ
nth	ɛnθ
nymph	nɪmf
warmth	wɔrmθ

Figure 4: Acoustic and Phonetic Classifications of the Word "length"

In summary, this paper presents a new approach to the problem of recognizing isolated words from large vocabularies and multiple speakers. The system initially classifies the acoustic signal into several broad manner categories. Once the set of potential word candidates has been significantly reduced through the utilization of the structural constraints, the acoustic differences between the remaining words can be examined in detail. Such a procedure will enable us to deal with the large vocabulary recognition problem in an efficient manner. What is even more important is the fact that such an approach bypasses the highly error-prone process of deriving a complete phonetic transcription from the acoustic signal. In this approach, detailed acoustic phonetic knowledge can be applied in a top-down verification mode, where the exact phonetic context can be specified.

REFERENCES

[1] Makhoul, J.I. (1975) Linear Prediction: A Tutorial Review, Proc. IEEE, 63:561-580.

[2] Itakura, F. (1975) Minimum Prediction Residual Principle Applied to Speech Recognition, IEEE Trans. Acoust., Speech, Signal Process. ASSP-23:67-72.

[3] Sakoe, H. and Chiba, S. (1971) A Dynamic Programming Optimization for Spoken Word Recognition, IEEE Trans. Acoust., Speech, Signal Process. ASSP-26:43-49.

7

[4] Keilin, W.J., Rabiner, L.R., Rosenberg, A.E., and Wilpon, J.G. (1981) Speaker Trained Isolated Word Recognition on a Large Vocabulary, J. Acoust. Soc. Am. 70:S60.

[5] Cole, R.A., Rudnicky, A.I., Zue, V.W., and Reddy, D.R. (1980) Speech as Patterns on Paper, in *Perception and Production of Fluent Speech*, Cole, R. A.(ed.) Hillsdale, New Jersey: Lawrence Erlbaum Assoc. pp.3-50.

[6] Shipman, D.W. and Zue, V.W. (1982) Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems, Conference Record, IEEE 1982 International Conference on Acoustics, Speech and Signal Processing. pp.546-549.

[7] Huttenlocher, D.P. and Zue, V.W. (1983) Phonotactic and Lexical Constraints in Speech Recognition. Proceedings of the 1983 American Association for Artificial Intelligence Conference, Washington, D.C. pp.172-176.

[8] Witkin, A.P. (1983) Scale-Space Filtering. Proceedings of the 1983 International Joint Conference on Artificial Intelligence, Karlsruhe, Germany, pp.1019-1022.

DISTRIBUTION LIST

	DODAAD Code	
Director Advanced Research Project Agency 1400 Wilson Boulevard Arlington, Virginia 22209 Attn: Program Manager	HX1241	(2)
Associate Director for Mathematical and Physical Sciences Research Program Office of Naval Research 800 North Quincy Street Arlington, Virginia 22217	N00014	(3)
Administrative Contracting Officer E19-628 Massachusetts Institute of Technology Cambridge, Massachusetts 02139		(1)
Director Naval Research Laboratory Washington, D.C. 20375 Attn: Code 2627	N00173	(6)
Defense Technical Information Center Bldg. 5, Cameron Station Alexandria, Virginia 22314	S47031	(12)
TACTEC Battelle Memorial Institute 505 King Avenue Columbus, Ohio 43201	79986	(1)